ABSTRACT

Title of dissertation: MEASURING LEARNING PROGRESSIONS USING BAYESIAN MODELING IN COMPLEX ASSESSMENTS

Daisy Rutstein, Doctor of Philosophy, 2012

Dissertation directed by: Professor Robert J. Mislevy
Department of Measurement, Statistics & Evaluation

This research examines issues regarding model estimation and robustness in the use of Bayesian Inference Networks (BINs) for measuring Learning Progressions (LPs). It provides background information on LPs and how they might be used in practice. Two simulation studies are performed, along with real data examples. The first study examines the case of using a BIN to measure one LP, while the items in the second study are designed to measure two LPs. For each study, data are generated under four alternative models, and each of the models is fit to the data. The results are compared in terms of fit, parameter recovery, and classification accuracy for individuals. In the case where one LP was used, two models provided high correct classification rates. When two LPs are being measured the classification rates were not found to be high, although an unconstrained model with freely-estimated conditional probabilities had slightly higher rates than a constrained model in which the conditional probabilities were given by lower-dimensional functions. Overall, while BIN show promise in modeling LPs, further research is needed to determine the conditions under which this modeling approach is appropriate.

MEASURING LEARNING PROGRESSIONS USING BAYESIAN MODELING IN COMPLEX ASSESSMENTS

Daisy Wise Rutstein

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2012

Advisory Committee:
Professor Robert J. Mislevy, Chair
Associate Professor Dan Chazan
Assistant Professor Hong Jiao
Assistant Professor Roy Levy
Assistant Professor Andre Rupp

© Copyright by
Daisy Wise Rutstein
2012

DEDICATION

To my husband, David, whose love and support has made this all possible.

ACKNOWLEDGEMENTS

There are many people without who I would not have been able to complete this program. I would first of all like to thank my family for all of their love and support. From helping me out with childcare to being a sounding board to offering encouragement they have been there for me.

I would like to thank the people at Cisco Systems, particularly Dr. John Behrens, Kristina Chapple and Dr. Kristen DiCerbo for the graduate assistantship opportunity. Working with this project was an excellent learning experience. I would also like to thank the Cisco team, including Patti West, for their help in providing me with the real-data used in this study.

I would also like to thank my fellow students who have given me much encouragement and been examples of the light at the end of the tunnel. In particular I would like to thank Dr. Heather Mann, Dr. Brandi Weiss, Dr. Marc Kroopnick, Dr. Will Dardick (who also provided me with computer support for this dissertation) and Dr. Roy Levy for their friendship and guidance throughout the years.

Most of all, I would like to thank my advisor, Dr. Robert J. Mislevy whose patience and support throughout this process has been above and beyond. His guidance has been invaluable in helping me develop not only as a student but also as a professional in this field.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: PURPOSE AND RATIONALE

The purpose of this study is to examine issues in modeling Learning Progressions (LPs) with Bayesian inference networks. While there are many definitions for learning progressions, the underlying concept is that they provide information regarding the state of a student in their level of understanding of a given concept. Learning progressions are broken down into levels, each of which should represent a given state of student's learning with descriptions of the types of knowledge and skills required for student's to display mastery of that level, or with descriptions of the types of concepts and ideas that students have at that level. The levels are generally considered to be ordered in the sense that higher levels indicate deeper levels of understanding of the given concept.

For both studying LPs and for inferences about students' progress in this light, we need to be able to create tasks that provide evidence about students' capabilities through the lens of the targeted LP(s), and have a statistical/measurement model for interpreting this data. This study focuses on the statistical modeling issues, in the framework of Bayesian inference networks (BINs). It concerns the question of how a BIN can be used to model the relationship between tasks which have observable responses and levels of the LP (which is a latent construct). Specifically, it will address the issues of recovery of the correct model as compared to similar incorrectly specified models, and the robustness of inferences about students from both the correctly and incorrectly specified models.

The study will first examine different models when all observables are modeled as depending on upon only one LP, and then it will address issues with how tasks can be modeled that depend on multiple LPs. In both of these cases several models will be proposed and compared through the use of simulation studies. This study will also provide a real data example to demonstrate how these concepts can be expanded and used in practice.

Learning Progressions: Background Information

Currently there is a movement in the measurement community to create assessments that can generate more cognitively and instructionally relevant information in addition to providing an overview of ability level of students (Leighton & Gierl, 2007, National Research Council (NRC), Nichols, Chipman & Brennan, 1995, 2001, Rupp & Templin, 2008). This information can be used to instruct learning in a classroom as well as to provide individualized information.

Some trace this movement back to the 1980's when there was a call for greater collaboration between cognitive psychologists and assessments developers, and greater collaboration between assessment and instruction (Huff & Goodman, 2007). This call was due in part to the fact that researchers have found that learning is optimized when there is an alignment among curriculum, assessment and some cognitive theory of learning (Huff & Goodman, 2007). In addition it has been found that assessment based on cognitive theory can have a positive influence on instruction and learning (Huff & Goodman, 2007).

It is not only researchers that believe in the benefit from diagnostic assessment, but also a high percentage of teachers think it is important to collect diagnostic information, whether it be from classroom assessments or large scale assessments (Huff & Goodman, 2007). Huff and Goodman (2007) also found that a large percentage of teachers wished they had more diagnostic information and, in particular, more individualized diagnostic information at the large scale assessment level of testing. Mislevy (1993) has also stated the importance of creating assessments that are able to provide meaningful information regarding students or classes of students.

With the introduction of No Child Left Behind (NCLB) there has been a recent increase in the articulation of standards, and interest in having students meet these standards (NRC, 2001). This in turn has increased the amount of testing (NRC, 2001). However, this testing generally does not give individualized information regarding the level of the student or information regarding methodology that will lead the students to meet the standards (Wilson & Scalise, 2006).

The NRC (2001) report stressed that formative and timely feedback is important to students in their development. If students are not given feedback in a timely manner, then they may continue to practice incorrect methodologies. The type of feedback that seems to be most beneficial is feedback regarding how the student is progressing towards the goal (versus feedback such as overall grade) (NRC, 2001). Formative diagnostic information can lead to guidance regarding what type of practice or instruction a student might need next, which can help dramatically in the improvement of a student's skill.

The National Research Council (2001) also states that while instructors often have set curriculum goals, they are responsible for any intermediate goals in the classroom. Having a theory of how to meet the curriculum goals based on knowledge of how students progress toward those goals can help in determining these intermediate goals, which in turn can influence how the curriculum is laid out.

One type of diagnostic information revolves around the use of learning progressions (Corcoran, Mosher, & Rogat, 2009). Learning progressions can be a useful tool when it comes to curriculum and assessment design. Concepts similar to learning progressions have been around for some time. For example, Piaget's Stages of Cognitive Development (Piaget, 1928) can be thought of as a learning progression for a student's ability to understand new material (Woolfolk, 2004). Gagne's work with learning hierarchies (Gagne, 1970) is another example of having a set of capabilities that have an ordered relationship to each other. However, learning progressions themselves are still being developed and there is need for further work in addressing issues such as the design, assessment, use, and modeling of these learning progressions (Wilson, 2009).

Learning Progressions: Definition

There are several ways of describing learning progressions. According to Popham (2007) learning progressions can be thought of as the building blocks for specific skills, or, put another way, the steps that one would take along the way to mastery of a task. This is similar to Gagne and Driscoll's (1988) concept of learning as a set of events that happen in sequence. From this point of view there are certain steps that must be taken in order for students to arrive at some end state. At the end of each step a student is in a given state and these states represent the learning progression.

Another way to define a learning progression is that it is a description of how students develop expertise over time (Stevens, Shin & Krajcik, 2009), which could incorporate the learning of new topics or gaining expertise from a basic level of facts to higher levels requiring more complex thinking. Learning progressions can be structured so that the lowest level is a novice level, or the lowest level could represent students who have a basic level of understanding. Similarly, Wilson (2009) presents a definition of learning progression as descriptions of how students change their thinking about a topic over time. White and Frederiksen (1990) discuss learning progressions as changes in the mental model. In their work with electricity they find that students start with a very naïve qualitative model of how things work. Their understanding progresses to incorporate more quantitative ideas and eventually they obtain expert models which incorporate both qualitative and quantitative concepts.

White and Frederiksen (1990) developed a progression of mental models for students understanding of circuits. These models incorporated different physical structures of the circuit, behavior of devices and basic electrical principals in order to demonstrate how students move from a low level of understanding to a high level. This can be further seen in Table 1.

Table 1: An example of a learning progression regarding circuits

Level	Learning Progression Levels
1	1) Understand that there are two polarities of electrical force, and both forces must be applied to two ports of the device 2) Understand that devices have properties, such as conductivity 3) Understand that devices can have more than one state which can determine the properties
2	1) Understand all Level 1 pieces 2) Understand series-parallel circuits 3) Understand the idea of a short 4) Refine their understanding of a conductive path into either a conductive-resistive path and a purely conductive path
3	1) Understand all Level 2 pieces 2) Understand and apply Kirchhoff's Voltage Law 3) Evaluate the effects of changes in conductivity on a device by device basis
4	1) Understand all Level 3 pieces 2) Evaluate the effects of changes in conductivity by propagation of voltages.

A similar concept of a learning progression can be seen in the Berkeley Evaluation and Assessment Research (BEAR) Assessment system (Draney, 2009, Wilson, 2009). Wilson (2009) has discussed how a learning progression can be built based on how students change their thinking over time. This includes not only how new knowledge is incorporated into a student's mental model but provides information about limitations in a student's understanding. These limitations can be misconceptions or areas in which the student does not have a clear picture (see Table 2 as an example).

Table 2: Detailed view of the Tracing Matter LP (taken from Draney, 2009)

Level	Accomplishment	Limitations
1	Macroscopic force-dynamic narratives about actors and events	Focus on reasons or causes for events rather than mechanisms Vitalistic explanations for events involving plants and animals
2	Stories involving hidden mechanisms Recognition of events at microscopic scale Tracing matter through most physical changes Coherent stories of food chains	Matter not clearly distinguished from conditions or forms of energy. Macroscopic events are associated with specific organs rather than cellular processes
3	Stories of events at atomic-molecular, macroscopic and large scales	Mass of gases not consistently recognized Incomplete understanding of chemical identities of substances
4	Model-based accounts of all carbon transforming processes	Difficulty with quantitative reasoning

In the recent report by the Center for Continuous Instructional Improvement (CCII), Corcoran, Mosher, and Rogat (2009) define a learning progression as a testable hypotheses regarding how a population of students' understanding and ability grows over time with appropriate instruction. In addition to this definition, the panel that was convened to discuss learning progressions in science for this report also came up with the following characteristics that every learning progression must have:

1) Learning targets or clear end points that are defined by societal aspirations and analysis of the central concepts and themes in a discipline;
2) Progress variables that identify the critical dimensions of understanding and skill that are being developed over time;
3) Levels of achievement or stages of progress that define significant intermediate steps in conceptual/skill development that most children might be expected to pass through on the path to attaining the desired proficiency;
4) Learning performances which are the operational definitions of what children's understanding and skills would look like at each of these stages of progress, and which provide the specifications for the development of assessments and activities which would locate where students are in their progress; and,
5) Assessments that measure student understanding of the key concepts or practices and can track their developmental progress over time. (Corcoran, Mosher, & Rogat, 2009)

This concept that learning progression should be based on research and testable is echoed in Stevens, Shin & Krajcik (2009). This requirement helps ensure that a learning progression is based on cognitive theory and requires evidence of the validity of the learning progression. If a learning progression is not able to be tested then there is no guarantee that having students follow the learning progression is an appropriate path.

The guidelines set by the CCII (Corcoran, Mosher, & Rogat, 2009) also directly tie the learning progression to curriculum and instruction by specifying activities that are most appropriate and by setting specific goals that can be reached. In addition there is a link between the curriculum and assessments in the requirement for assessments to be developed. This again helps to provide evidence for the validity of the assessment, when used to make inferences about the student's ability on the learning progression, given that the learning progression itself has been validated.

Behind any of these definitions of learning progressions is the concept that there are different stages students go through when obtaining a deep understanding of a subject. Determining which stage a student is in can not only help determine what skills they have mastered, but also what steps they should take in order to progress to the next stage. An instructor with information about the stage a student is in can then determine what they need to cover in their instruction to best help their students.

A distinction that is crucial to this dissertation arises at this point. It is the difference between a learning progression and variables in statistical models that may be used to organize reasoning with evidence and uncertainty about individuals and groups with respect to performances and learning progressions. Learning progressions, as they have been described in this section, are psychological schemas for the nature of cognitive development and its manifestation in task performance. Measurement models are statistical overlays on top of the substantive psychological theory, for rigorous handling of evidence. A key point is that there is no simple unique relationship between a psychological learning progression conception in general and a universal measurement model.

It is clear from the previous discussion that there are variations of learning progressions as a psychological concept, and none are defined specifically enough to uniquely pinpoint the form and parameters of a specific measurement model to accompany it. This is an applied engineering problem: Given a particular learning progression model and tasks and performances meant to provide evidence about it, alternative measurement models could be entertained. Thus, when we speak of a learning progression, we should ideally indicate that we are speaking of the psychological conception, and when we speak of learning progressions variables, we should indicate that we are speaking of the formal variables in a statistical model that are representing some aspects of students' capabilities in the psychological model.

For example, the CCII report (Corcoran, Mosher, & Rogat, 2009) differentiates between the concept of a progress variable and a learning progression (also see Wilson, 2009). Progress variables are defined as variables that define growth points for students along the scale in a measurement model (Wilson & Scalise, 2006). , Both incorporate the idea that students progress from low level attributes to high level attributes (NRC, 2001), although the progress variable is an instantiation of a particular modeling approach and definitions and procedures within it. A learning progression may be made up of several progress variables, and the relationships among these variables could be complex (Wilson, 2009).

Learning Progressions: Background Research

When it comes to curriculum development, learning is not a straightforward march through a series of steps, but rather a dynamic path full of leaps forwards and setbacks (Corrigan et al., 2009). Learning progressions can pinpoint landmarks in a students learning, and the development of the learning progression can provide information about what type of instruction would be best at these different stages (Corrigan et al., 2009).

Research that has been done in the field of expert-novice research (e.g., Ericsson et al., 2006) may be helpful in defining a learning progression. This work examines differences between experts and novices, and helps to indicate some of the key attributes that should be included at different levels of the learning progression. Included in these findings is that a key difference is knowledge organization (NRC, 2001). One example of this is a study by Chi, Feltovich, and Glaser (1981) in the field of physics. Here students were asked to place particular problems into different groups. Novices tended to group these problems by what type of device they were using (such as pulleys vs. planes). While experts tended to group based on the underlying principles (such as Newton's first laws vs. conservation of energy) (NRC, 2001). A learning progression for physics may then have at a low level that students are able to recognize similar problems by physical objects, while at a higher level students are able to recognize the principles of a problem. This is something that is testable, as it is conceivable to develop a problem that tests how students organize their knowledge and may be a useful way to determine the level of a student. While this is one area of research that may reflect on the development of learning progressions further research into the development process is needed.

One issue with developing learning progressions is that often, not all students follow the same learning path (NRC, 2001, Stevens, Shin & Krajcik, 2009). The concepts of learning paths will be discussed in more detail in Chapter 2 but it is important to note that in some cases it may be difficult to find a strict progression that students are expected to follow. The different paths a student can take should be taken into account when developing the model for the learning progression (NRC, 2001). In general the learning progression should cover the general pattern of learning, with concrete differences between the different levels of the learning progression, and should include information regarding how to help students progress through the levels, as well as how to measure the level of a student. (Stevens, Shin & Krajcik, 2009).

For the purpose of this study, the term attributes will be used to describe the pieces that make up the layers of the learning progression. As mentioned above the learning progression may give descriptions of knowledge a student has or skills the student should be able to display, but it may also contain information regarding misconceptions and frameworks that a student might have. This study is not examining the specific pieces of individual learning progressions and therefore attributes will be used as a general term to describe what a typical student at a given level may look like.

Assessment Triangle and Evidence Centered Design

One area of current research revolves around the generation of assessments that provide diagnostic information (Wilson & Scalise, 2006, Gotwals, Songer & Bullard, 2009). This literature addresses two sides of the story. One is from the development point of view, where the question is how an assessment can be created in order to measure the attributes associated with the learning progression. The other is from an analysis point of view, where one would determine how to analyze the assessment in order to obtain the desired information regarding the student’s level of ability.

When developing an assessment both of these questions should be considered jointly as it is important to determine how an assessment will be analyzed when it is created, and keeping the purpose in mind will help determine how an assessment should be analyzed. The National Research Council (2001) defines three elements: cognition, observation and interpretation that make up what is referred to as an assessment triangle (see Figure 1). These elements must work together in order to create valid assessments. Cognition is defined to be the theory of learning, or what it is we want to say about the student. Observation is the kinds of tasks that allow the student to display information regarding what it is we want to measure. Interpretation is the link between the observations and cognition, or how information about the attributes of the task can provide information on the beliefs regarding the student (NRC, 2001).

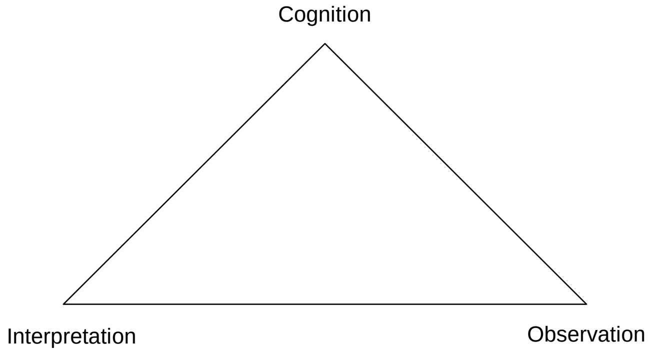


Figure 1: A representation of the Assessment Triangle

These pieces of the assessment triangle are also developed using an evidence-centered design (ECD) approach to creating assessments (Mislevy, Almond & Lukas, 2003). An ECD approach starts with the domain analysis stage, where information is gathered regarding the domain in question, moves to the domain modeling stage, where this information is organized, and then moves into the conceptual assessment framework (CAF) stage (Mislevy, Almond & Lukas, 2003). It is at this stage that the three main models, the student model (what it is we want to say about the student), the task model (what type of tasks would allow the student to exhibit the behavior), and the evidence model (how we can use the information from the work products produced by the task model to make inferences regarding the student model) are developed (Mislevy, Almond & Lukas, 2003).

Notice that these three models are very similar to those in the assessment triangle. The difference is that the models in the CAF layer of evidence-centered design are formal syntactic models for the operational elements of an assessment, as opposed to the psychological concepts that make up a substantive assessment argument. For example, in a student model to be used in conjunction with a learning progression, the learning progression itself is the substantive and psychological theory of the increasing states of knowledge, and the student model consists of latent variables in a psychometric model that are used to represent students' standing within the frame of the psychological theory.

When developing an assessment, it is very important that these different models are coordinated. Using the structures presented in the ECD framework helps ensure the validity of the assessment, as the reasoning for each of the decisions made for the elements of the assessment are laid out and the backing needed to support those decisions is explicit (Mislevy & Riconscente, 2006).

A learning progression should have specific targets in mind for each level, which provide information regarding the student model for tasks designed at each level of the learning progression. One learning progression may lead to different student models but the information needed to determine what these models are should be provided. In addition, the learning progression, as specified by the CCII report (Corcoran, Mosher, Rogat, 2009), should provide information regarding the type of tasks that can give insight into the student model, i.e. it should provide information regarding the task model. Again many tasks may be developed to measure a given learning progression, but the information needed to create these tasks should be provided in the learning progression.

As noted above, there is a natural relationship between the theory of increasing capabilities in a learning progression with variables in a psychometric student model, and the former is the center of discussion in research on learning progressions. Less explicit, however, from discussions of learning progressions is information regarding the evidence model, or how one uses the observations provided by the task to provide evidence with regard to the learning progression. In part this is because there are many different methods that can be used to formalize and operationalize the notion of learning progressions in psychometric models, and different models may be appropriate for different situations. This research is targeted at examining one small piece in the area of the evidence model.

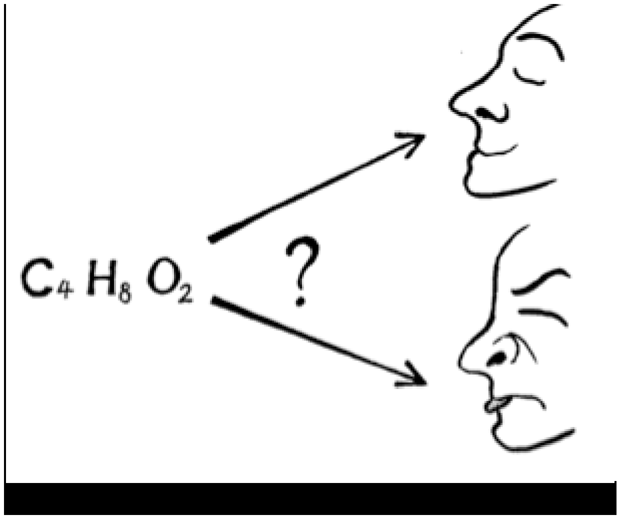
Learning Progressions and Assessments

As mentioned above, a learning progression should provide insight into how an assessment can be structured. The learning progression should have specific goals for each of its levels. These goals can then be used as the student model for an assessment. For example, using the learning progression for circuits discussed above, several different student models may be conceptualized. One student model may be regarding students' ability to explain conductivity; while another may be that students can apply Kirchhoff's Voltage Law.

Once the student model is determined then tasks can be developed that would measure the attributes specified by the student model. The CCII report (Corcoran, Mosher, Rogat, 2009) states that information regarding tasks that can be used to measure the different levels of the learning progression should be included in the learning progression. This information can be used to help develop assessment tasks. In some cases, these tasks would reflect on one level of the learning progression. For the learning progression on circuits, if the student model is one in which the student is able to explain conductivity, then since this attribute is included at Level 1 of the learning progression, the task designed to measure that attribute would be designed to help determine if the student has one of the attributes required to be at Level 1.

Tasks can also be developed that would be geared towards multiple levels of the learning progression. For example, in the matter learning progression described above (see Table 2) a student model could be: the student is able to explain the relationship between molecular formulas and structural formulas. This is an attribute that runs across multiple levels of the learning progression, as the higher the level of the learning progression, the higher the students' ability is in this area. The decision must be made when creating the task whether the task should provide the opportunity for students to answer at different levels (in which case the task could be used to determine the student's level) or if it just allows for responses that are at a given level (in which case the task would only determine if the student has the attribute appropriate for that level or not).

For example, take the following task:



"Both of the solutions have the same molecular formulas but butyric acid smells bad and putrid while ethyl acetate smells good and sweet. Explain why these two solutions smell differently." (Draney, 2009)

Figure 2: Sample task based on the matter learning progression

For this task students are able to respond freely. Students who are at Level 1 of the learning progression may give a reason such as maybe one of the solutions went bad or is older which doesn’t use molecular chemistry concepts (Draney, 2009). While students at level 2 may state the fact that they might have different structural formulas but not go into details regarding this difference (Draney, 2010).

In contrast an item such as a true/false item that states “True or False: if two solutions have the same molecular formula then they must also have the same structural formula” would be aimed at providing evidence on Level 2 of the learning progression, and does not have the opportunity for a student to display higher level attributes. Following this idea, tasks can be characterized as to what levels of the LP they can discriminate between.

It is up to the test developer to determine which type of task is more appropriate for the given assessment. The developer may want to target the entire assessment at a particular level of the learning progression or they may want to use the assessment to determine a student’s level on the learning progression. When using the assessment to determine a student’s level, both types of tasks are appropriate, as the assessment could contain several different tasks that are designed to measure different levels of the learning progression.

What will help in determining what items should be used is the evidence model, and in particular information regarding how the different pieces of evidence will be accumulated to reflect on the student model. The evidence model is a key step in ensuring that the information gained from the tasks reflects accurately back onto the student model. Currently the definition of a learning progression does not have information regarding appropriate evidence models, and for either a research program to investigate and refine a particular learning progression or an operational assessment to character students’ standing with respect to the progression, it is up to the test developer to determine how evidence is acquired and accumulated.

For this research, the items are assumed to be items that are designed to measure specific levels of the learning progression. In order to have evidence regarding the different levels of the learning progression, items that are targeted at the different levels are combined into one assessment. The next section will discuss different evidence models.

Modeling Learning Progressions

Standard measurement practices for assessments include developing the overall construct as a continuous unobserved variable and then creating the observables as categorical variables with numbers assigned to them (NRC, 2001). For example, 0 and 1 for incorrect and correct answers on dichotomous items, or a score between 0 and 4 on a rating scale for an open-ended performance. Common ways of modeling performance on such tasks make use of classical test theory (CTT) or item response theory (IRT) (Hambleton & Swaminathan, 1985). These methods are generally more applied to a summative type of feedback (versus formative), as they tend to give an overall summary of the student’s ability and not information about specific strengths and weaknesses. Generally in testing, particularly large scale testing, the concern is with the location of a person along the overall proficiency scale, or on specific subscores, or how much of some ability a subject has, instead of the cognitive background regarding why a student is at that location (Leighton & Gierl, 2007).

A more recent trend in modeling has been the development of cognitive diagnosis models (CDMs) (Rupp & Templin, 2008, Rupp, Templin & Henson, 2010). These are models that are used to connect categorical observable variables with latent classes with the ability to provide formative feedback (Rupp & Templin, 2008). These types of models have been referred to by many different names in this growing area of research. Further discussion of these models will be given in Chapter 2, but it should be noted that these models seem to be more appropriate for modeling learning progressions than traditional CTT and IRT methods, as CDMs are designed to provide diagnostic information instead of information on ability level.

West et al. (2009) proposed the use of Bayesian inferences networks (BINs), a general modeling framework in which CDMs can be instantiated, to model data from tasks meant to evidence students' status on learning progressions. BINs have been applied in educational assessment as a particular class of psychometric models (Almond, Dibello, Moulder, & Zapata-Rivera, 2007) with latent student-model variables that represent aspects of students' knowledge or skills, to determine probability distributions for the values of observable variables derived from students' task performances. Although CDMs is itself a general approach that can be implemented in various ways, the use of BINs for this purpose is motivated by the advantages noted below.

This research will build on the work by West et al (2009) and examine in more detail the use of BINs. While other methods have their own strengths, and further research can be performed to compare different methodologies, BINs have advantages that are of interest. For instance, once the network has been set up, inferences can be drawn based on partial data. BINs are very flexible, in that they can handle many different types of models and different types of observable variables (Mislevy, 1994, Schum, 1994). As West et al (2009) mention, BINs have been used in educational settings and allow for the user to model the structure of the variables as well as the nature of the probabilistic relationship between variables. In addition BINs have the flexibility to be extended or concatenated when more elements are brought into the modeling problem, such as multiple learning progressions, new tasks, and additional observable variables from existing tasks. The BIN framework can be used to instantiate other CDMs, and constraints can be placed onto the structure and conditional probabilities of a BIN so that it can provide discrete approximations of classical test theory and item response theory models.

Multiple Learning Progressions

Students generally do not learn just one skill at a time. Often these skills are related skills (e.g., prerequisites), and in some cases it may be hard to assess one skill without using tasks that also rely on another skill. Thus it is important when developing a learning progression to also think about the relationships between different learning progressions (Corrigan et al., 2009), and when tasks involve multiple skills, the ways in which performance depends on those skills. This thinking can be used to help improve curriculum and instruction, and must be taken into account when developing assessments.

While most of the recent work in the context of learning progressions has dealt with a single learning progression, the question of how to model multiple learning progressions is an important issue. The development of multiple learning progressions may occur at different time points and the relationship between them may not be made clear. While some LPs may surround skills that are not related, others may be directly related and others may have more complicated relationships.

There is also the issue of how one assesses the learning progressions. This again can be addressed from the assessment design point of view as well as the modeling point of view. When it comes to design, the decision must be made regarding whether or not tasks are designed to measure multiple LPs or just one. Some of this may depend on the relationship between the different skills as well as any constraints on the assessment. Based on how these tasks are designed different models may be used to analyze the assessment. The choice of an appropriate model is important when it comes to the validity of the assessment.

The work by West et al (2009) only addressed the issue of one learning progression. However, a BIN can be expanded to incorporate multiple learning progressions. This adds new levels of complexity to the model, and questions such as how the learning progressions relate to each other, and how they relate to the observables when the observables are designed to provide evidence about both learning progressions must be addressed.

Study Purpose and Overview

There are many choices for how to model data in order to obtain diagnostic information regarding students. The choice regarding which model to use may depend on how the learning progression (if used) is set up, and could in fact influence the development of the learning progression. One possibility for model choice is the use of a BIN. In order to implement a BIN, decisions must be made regarding how to set up the network and how to model the relationship between the LP (or LPs) and the observable variables.

This research will provide insight for some of these choices by exploring modeling options from the Bayesian network and cognitive diagnosis literature. These options will be used to develop alternative BIN models, which will then be examined with respect to parameter recovery and robustness of inferences regarding individual students. Of particular interest are implications for model choice, such as whether certain models are sufficiently robust to justify their use, even in cases when they may be misspecified. The information from this research should help a practitioner who is using BIN for modeling learning progressions make appropriate model choices.

Alternative models expressed in the BIN framework will be presented that will represent the relationship between observable variables and student model variables. This relationship is expressed through conditional probability distributions, specifically, probability distributions for possible outcomes on the observable variable, given values on the student model variable(s) posited to determine performance on the task. The models chosen will highlight how different decisions may be reflected in the model. The research will examine how different constraints on the relationship between observable variables and their corresponding learning progressions affect parameter recovery in estimation and the robustness of inferences from the model. It will be assumed that the learning progressions are well defined and tasks are targeted at particular levels of these learning progressions.

Two kinds of constraints will be made on the conditional probabilities: (1) constraining them in a manner that reflects the hypothesis about the relationship between the learning progression and the observable variables and (2) using the latent class Rasch model (Formann & Kohlmann, 1998) to approximate the unconstrained table of conditional probabilities. A three part study is proposed:

1. Study 1 will focus on the case where observable variables depend on only one learning progression. There are a number of related but distinct BIN structures that are consistent with the general concept of a single learning progression. The question addressed here is: Are there circumstances in which it is beneficial, for purposes of classification, to model a learning progression in terms of latent variables for the levels of the proposed progress, as opposed to its one latent variable? The study will compare a model with the learning progression represented as one categorical latent variable, and models in which the attributes of the learning progression are treated as separate categorical variables with varying hierarchical constraints amongst these variables. The comparison will examine overall classification accuracy for different populations as well as parameter recovery and model fit.
2. Study 2 will address the case where observable variables depend on two learning progressions (i.e., at least some observable variables have two student model "parents," both of which embody a learning progression). This research will address the question of whether or not putting constraints onto the relationship between the two learning progressions and the observable variables improves classification accuracy of the students. Three different constraints, namely compensatory, conjunctive and disjunctive, will be taken into consideration along with an unconstrained model. Again the models will be compared in terms of classification accuracy, parameter recovery and model fit.

3. In Study 3 two real data examples will be presented that will demonstrate the implications of using the above mentioned models in practice. Comparisons of model fit as well as differences in the conclusions drawn from the application of the hypothesized relationships will be discussed, along with practical issues that may arise when using a BIN in practice.

The first two studies, then, are simulation studies that generate data using different constraints under similar models, and compare the results to determine the performance of the individual models. The results of these studies should give practitioners some insight into the consequences of different decisions that must be made when using a BIN. In addition, the real data example will provide concrete information into how different BIN models can be used in practice. The combination of the studies will highlight decisions that need to be addressed and the appropriateness of particular models.

CHAPTER 2: LITERATURE REVIEW

While learning progressions are relatively new and the literature is still being developed, there is relevant research in related fields. There are results both on the development side of learning progressions where concepts from learning paths can be beneficial when determining how students progress, and on the modeling side where latent class analysis and cognitive diagnostic modeling research can be applied.

A learning progression can be represented by a categorical latent variable. An underlying concept that is being measured but cannot be observed directly; therefore it is latent. The learning progression consists of (usually ordered) levels, such that students can be at any particular level of the learning progression. These levels may consist of information from different progress variables but there is still a clear distinction between the given layers. Since there are a finite number of levels, the variable is categorical. When observable variables in the form of evaluations of aspects of students' performances (e.g., item responses, ratings of efficiency) are also categorical, research in latent class analysis is directly relevant.

Also relevant is the field of cognitive diagnostic modeling (as discussed in Chapter 1). This field investigates how to measure latent categorical variables in order to obtain diagnostic information regarding aspects of students' knowledge and skills and therefore is also relevant. Bayesian inference networks (BINs) described in Mislevy (1994) are another type of modeling approach that can be adapted to provide diagnostic information, and can be particularly useful when multiple attributes are being measured.

This chapter will discuss learning paths and modeling techniques in relation to learning progressions. The following section describes learning paths as they have been studied in science and mathematics. The subsequent sections draw on research in latent class analysis and cognitive diagnosis models to discuss modeling students' movement through such paths and observable evidence of this movement.

Learning Paths

One decision that needs to be made when determining a learning progression is what constitutes the different levels of the learning progressions. In general, the higher levels of the learning progression should correspond to higher capabilities, whether these are higher order thinking skills or attributes that build on the lower level attributes. The term learning progression also implies that students would obtain the lower level attributes first and then progress through the different levels (although there may be situations where this is not the case, as will be discussed below.) Determining the relationship between the attributes required at each level of the learning progression can provide insight into how this relationship should be modeled.

The research carried out in this dissertation addresses performance at a single time-point; that is, it concerns cross-sectional rather than longitudinal observations. Such data can provide insights into the structure of variables and variable states to describe a learning progression, and conditional probabilities of task performance given states. Cross-sectional data cannot, in and of themselves, provide direct evidence about the paths that students take through a learning progression. For completeness, this section briefly notes work on learning paths and learning trajectories, because it has been associated with learning progressions in the literature.

Stevens, Shin and Krajcik (2009) describe a learning trajectory as a subset of a learning progression in the sense that it addresses a specific learning goal, but additionally includes information regarding how students can meet that learning goal. This information may include possible difficulties for the student and different misconceptions the students may have. A learning progression then is a collection of learning goals which can be attained through one or more learning trajectories.

When developing a learning trajectory, an analyst can take different learning paths into account. The learning trajectory is developed from a path that is deemed most appropriate (often based on research) and a model is built from that path. Examining the learning path that is chosen can then determine the relationship between different skills/abilities associated with the specific learning trajectory or learning progression. Research in the field of learning paths can help to determine how skills and abilities may be modeled in the learning progression. In particular, key concepts or skills that re-appear in certain sequences across multiple learning paths are candidates for stages of a learning progression.

For example, Mohan and Anderson (2009) generated a learning progression for the carbon cycle by first creating a framework in which this learning took place, and then developing an understanding of typical paths that students took when going from a low level of understanding to a high level of understanding. From these paths they were able to develop their learning progression. In their work, Mohan and Anderson (2009) found that students start by developing a language to discuss the events they see in nature. The progression to Level 2 involves the student’s ability to recognize hidden mechanisms, or constructs that are not seen by the human eye, as causes for certain events. It was found that there were two key cycles which when learned helped in the transition from Level 1 to Level 2, although the ordering of learning these cycles was not noted as not important (Mohan & Anderson, 2009).

The transition from Level 2 to Level 3 involves the recognition that matter is transformed. However, at this level students still do not have the sophistication in understanding chemical substances and the use of energy which can be seen in students in Level 4. Again, students may take different paths, by learning about different subjects or learning different concepts in different orders to transition between these states, but in general it was found that these are the stepping stones for students (see Figure 3)

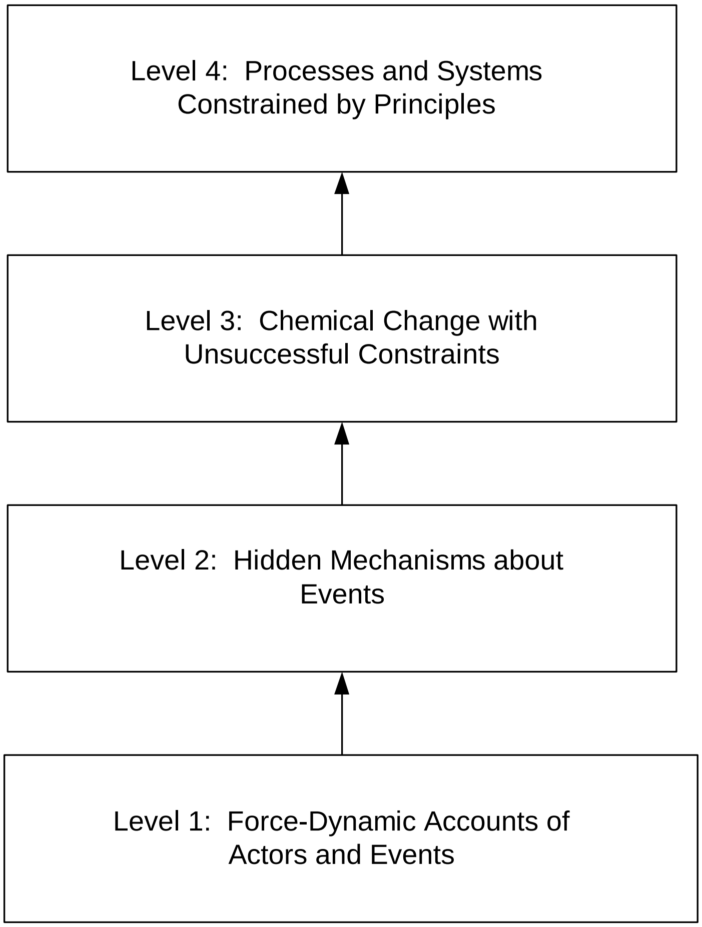

Level 4: Processes and Systems Constrained by Principles
Level 3: Chemical Change with Unsuccessful Constraints
Level 2: Hidden Mechanisms about Events
Level 1: Force-Dynamic Accounts of Actors and Events

Figure 3: Students' movement through learning with regards to the carbon cycle (Taken from Mohan and Anderson, 2009)

Haertel and Wiley (1993) discuss the acquisition of skills in reference to the creation of learning paths. They describe the simple case where a particular skill that is being learned can be broken down into two subskills. There are two main learning paths that the student could take in order to master the main skill. The main difference between these two paths is the relationship between the subskills. In one case, one skill is a logical prerequisite for the other skill (see Figure 4). In the other case, learning of the skills could be only partially ordered (see Figure 5), in which case the student may learn either skill before the other (although the skills could still be statistically dependent; for example, although both (A, ~B) and (~A, B) can occur, (A, ~B) may be much more frequent). These might be described as "hard" and "soft" prerequisition relationships, which would then be modeled differently.

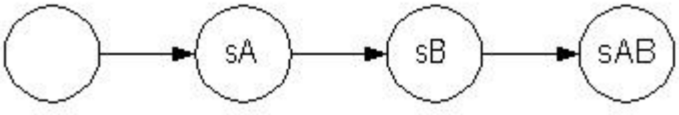


Figure 4: Learning path where skill A is a prerequisite for skill B (Haertel & Wiley, 1993)

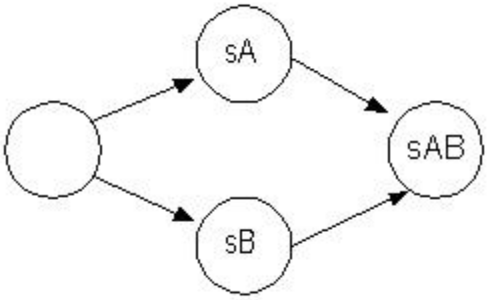


Figure 5: Learning path where skills are independent of each other (Haertel & Wiley, 1993)

A learning trajectory or progression can build from a learning path by examining the different steps that students may take and breaking those steps into different levels. At any given level there are specific attributes that the students would have. In a learning progression a person at a high level is generally believed to have not only the lower level attributes but to also have some additional attributes. These attributes could build on lower level attributes, such as going from a lower level of understanding to a higher level of understanding, or could be separate attributes.

Consider for example a learning progression for addition. While different researchers could theorize this progression in different ways, for the purpose of demonstrating features of learning progressions here addition will be modeled as consisting of four attributes:

Attribute 1: Ability to recognize the problem as an addition problem

Attribute 2: Ability to add two 1 digit numbers

Attribute 3: Ability to carry

Attribute 4: Ability to add two multi-digit numbers

Generally it may be believed that students would progress through the levels as they are laid out. In this sense the learning trajectory could be broken down into 4 levels, with each level indicating that the student has obtained the attributes corresponding to the level number and all of the previous attributes. However it may be the case that a student may be able to perform a carry operation before they have actually learned how to add two single digit numbers. If this were the case then a different learning trajectory could be generated in which students at Level 2 would be able to recognize an addition problem and perform a carry operation and then Level 3 would correspond to being able to recognize an addition problem, perform a carry operation, and add two one digit numbers.

The decision must be made which learning trajectory is more appropriate and then the learning progression will be built with that trajectory in mind. However, when modeling the learning progression it must also be determined whether there is room for multiple paths or whether the relationship between the two skills should be kept as a strict hierarchy. The learning progression could follow one of these learning trajectories, or if both were fairly common perhaps combine the two middle levels into one level, or define the second level as having either attribute 2 or attribute 3 (the first level still simply requiring the students to be able to recognize an addition problem) and then the third level would require the students have both attributes.   While this last approach precludes being able to distinguish between the two patterns that constitute the middle level of the progression, the resulting model may be more useful when classifying students.

Another possible way to define the levels of a learning progression is by determining different misconceptions a student may have (Wilson & Scalise, 2006). In some areas, such as science, there are general misconceptions that students seem to have at various stages. If the levels represent how the students move through those misconceptions, then being at a higher level doesn’t necessarily mean that the student has mastered the attributes at the level below, more that they have moved their understanding past that level. In this case, while the learning path may be linear through the different levels, the relationship between the levels is not so linear. Students could easily jump over a common misconception, therefore skipping a level in their understanding. This again must be taken into account when designing the model as now the underlying attributes aren’t related to each other per se.

In terms of learning trajectories and learning progressions, while the trajectory may specify how students normally transfer between the different states, and may give information regarding tasks and learning processes that will help the students transfer, the learning progression may be thought of as a state machine. In this sense students are in one state at a time, and while over time there may be a relationship between how they transition, at any one point in time a student can only be in one state, and the probability of being in another state given they are in a first state would be zero. The learning progression may then be represented as a categorical latent variable where the different categories represent each of the different states.

Once the relationship between the different levels of the learning progression is determined, and any attributes that are part of this learning progression are defined then the question revolves around determining how to model the relationship between the latent variables and the observable variables. This relationship will differ depending on the relationship of the levels of the learning progression. In addition, the types of tasks needed to provide information to reflect on the learning progression may differ depending on the type of learning progression. When determining how to measure the relationship between the learning progression and the observable variables, information from latent class analysis and cognitive diagnosis modeling can be applied.

Methods to Obtain Diagnostic Information

As mentioned in chapter 1 typical methods to model data from an assessment include using classical test theory (CTT) or item response theory (IRT). While generally this information has been used to determine where a student is along a given ability scale, there are expansions to this work that are geared towards obtaining diagnostic information.

One method is to make use of subscores. In this instance, an assessment would be developed in which different items reflect upon different attributes. This can be as simple as adding up the points for each item that reflects upon the given attribute. Or more complex methods can be used by using linear combinations of items (Haberman & Sinharay, 2010)

A more recent approach has been to use multidimensional item response theory (MIRT) models (Reckase, 2009). These models are extensions of the standard IRT models, but instead of an estimation for the ability on one attribute, ability parameters on multiple attributes are estimated (Haberman & Sinharay, 2010).

While these methods are useful for measuring levels of multiple abilities, in the case of learning progressions the attributes are generally related and the question being asked isn't regarding a students' ability on several different attributes but rather where the student lies along a single learning progression variable. For this type of information latent class models are more appropriate.

Latent Class Analysis

Latent class analysis (LCA) provides a methodology for modeling a categorical latent ability based on categorical observable data (McCutcheon, 1987). In terms of learning progressions it is the methodology by which the level of the learning progression can be determined for particular students based on their responses to observables. (The observables must be categorical. Analogous techniques exist for situations in which the latent variables are categorical and the observable variables are continuous but they fall under the realm of latent profile analysis; McCutcheon, 1987).

Latent class analysis is one method used to model the relationship between observable variables. The general formula associated with latent class analysis is given by: $\pi_{ij..mt}^{AB...EX} = \pi_{it}^{AX}\pi_{jt}^{BX}...\pi_{mt}^{EX}\pi_{t}^{X}$ where A,B…E are categorical variables that are dependent on the latent variable X and i,j,…m,t are states corresponding to those variables and π is the probability associated with being in the given states (McCutcheon, 1987). This formulation assumes that the observable responses are locally independent (as latent variable psychometric models generally do). In other words, that the probability of responses on the different items depends on one or more additional variables; conditional on the values of these variables, the responses are independent. For example, in a questionnaire with different questions about the government's responsibility when it comes to the environment, the responses may depend on a person's political affiliation. Once this affiliation is known the probabilities of the individual questions responses are determined and are statistically independent. Learning the response to one of the questions does not change the belief regarding the probability of the responses to other questions.

In latent class analysis, there is an overall ability (or attitude) that is assumed to exist but be latent (cannot be observed directly), it renders observed responses independent, and it is the overarching attitude or ability that the questions are designed to measure. The latent variable is also assumed to be categorical and people are assumed to fall into one category (although there have been studies done with regard to what to do with people who can not be categorized. This will be discussed briefly in Chapter 4).

The general methodology behind latent class analysis is that data is collected regarding subject's responses to the observables and (for at least a sample of people) estimates of the conditional probabilities are found by marginalizing over the probabilities that each subject is in each of the classes. This data is then used to estimate the probability of class membership for the latent class variables and the probability of responses for each of the observables given the latent class membership.

Latent class analysis can be used in both a confirmatory and an exploratory approach. In an exploratory setting, probabilities are estimated for several competing models which differ by the number of classes they contain. The best fitting model, in terms of most probable model, is then selected (using for example, the likelihood ratio or a modified version of it such as AIC or BIC that takes sample size and/or number of parameters being estimated into account; see Burham & Anderson, 2004). In a confirmatory approach different constraints can be placed on the probabilities and these constraints can be tested. For this type of analysis the constraints can be tested by examining the overall fit of the model with these constraints in place. In the context of learning progressions, the use of LCA will tend to be confirmatory because theory about the learning domain, how students move through it, and how their capabilities are evidenced in certain kinds of performance in certain kinds of tasks provides a strong initial hypothesis for the structure of the relationships of the variables. Exploratory uses of LCA are more suited to very early stages in one approach to defining learning progressions, namely exploring patterns of responses to existing assessments to determine whether patterns that signal underlying learning progressions may be present in the data.

The constraints in latent class models may be implemented by setting certain probabilities equal to a given value, setting equality constraints, setting inequality constraints, or modeling certain conditional probabilities in terms of parametric forms with fewer parameters. For example, it may be posited that members of a given class do not ever respond in a particular manner, in which case the conditional probability of that type of response for members of that class can be set to zero. One other type of constraint is used in latent class scaling analysis, in which the probability of certain responses must increase (or decrease) for a certain ordering of the class membership. Even within this type of analysis further constraints may be made such as setting error probabilities (the probability of answering in a manner not consistent with the given class membership) equal.

Since a learning progression can be represented as a categorical latent variable the methodology in latent class analysis can be directly applied. However, within latent class analysis there is still a large choice of models that can be used. The field of cognitive diagnostic modeling has taken many of these concepts developed in latent class analysis and applied them to the development of models that can provide diagnostic information. This type of information can be used to further help in identifying the attributes that students have and therefore the students' level along the learning progression. Bayesian inference networks (BINs) are one type of cognitive diagnostic model that is particularly well suited for modeling learning progressions. An overview of cognitive diagnostic modeling will be provided in the next section, followed by specific information regarding a BIN.

Cognitive Diagnosis Modeling

There has been much recent development of cognitive diagnosis models (CDMs). These models have been studied under many labels (such as cognitively diagnostic models, cognitive psychometric models, latent response models and structured located latent class models (Rupp & Templin, 2008)), but the use of the models to obtain information regarding diagnostic feedback remains the same. One central concept regarding the development of these types of models is that they should be tied to theory developed from a cognitive psychology viewpoint (Rupp & Templin, 2008). This theory is involved in determining what variables are best for the model at hand and the relationship between these variables. Theory can also help determine the relationship between the observable variables and the latent variables.

CDM's are a type of latent variable modeling in that they involve latent variables to be modeled. These variables are generally the skills required by the assessment. However, they differ from the traditional univariate view of latent variable models used in large-scale testing (classical test theory and item response theory) in that they contain multiple latent variables (Rupp & Templin, 2008). Note that in the equation for latent class analysis there was one latent variable state, X which corresponds to one latent trait. In a CDM the latent class c is defined by the students' ability on multiple attributes.

While the number of latent variables and the hierarchical structure imposed on these variables in a CDM may vary for a diagnostic model, there is generally more than one attribute of interest and the model should help determine the set of attributes obtained by the student. In the case of a learning progression, there may be one overarching attribute but this can still be broken down into different attributes, as each level may represent different sets of attributes. In this sense, the information obtained from the model is not simply what class a student is in, but instead what are the student's attributes. The decision of whether to break out attributes into variables, and if so how, is mainly an issue of grainsize. Some of this may depend on the specificity of the conclusions that are to be drawn from the assessment, and whether or not those conclusions are about the specific attribute levels or a higher more general attribute level. In addition CDMs often also have complex loading structure (Rupp & Templin, 2008), since tasks may depend on multiple dimensions—that is, on some combination of attributes.

Another difference between CDMs and general latent class models is that the type of constraints that are placed on the model fix the number of latent classes to be estimated (in addition to, as stated above, defining each class by the attributes required of a member of that class) (Templin & Henson, 2006). Therefore, the model is used from a confirmatory approach and is not generally used as an exploratory tool.

This loading structure is often represented by a Q matrix. This is a matrix that indicates for every item which attributes it requires. For the addition example shown earlier in this chapter, an exam could be created that has 8 items, 2 items designed to measure each of the attributes (See Table 3). Note that in this example, items that measure a particular attribute, also require the previous attributes. This matrix can not only help with the analysis of the exam, as it is clear which attributes the items are designed to measure, but also in the creation of items for the assessment, as this type of information makes some of the requirements for each item clear.

Table 3: An example Q-matrix for an exam with 8 items depending on 4 attributes

Item Number	Attributes			
	1	2	3	4
1	1	0	0	0
2	1	0	0	0
3	1	1	0	0
4	1	1	0	0
5	1	1	1	0
6	1	1	1	0
7	1	1	1	1
8	1	1	1	1

The general model for a CDM follows from a latent class model (for binary outcome variables) and is as follows:

$P(X_r = x_r) = \sum_C v_c \prod_I \pi_{ic}^{x_{ir}} (1-\pi_{ic})^{1-x_{ir}}$ (Rupp, Templin, & Henson, 2010) where x_r is the vector of response data for person r (responses are assumed to be binary), v_c is the probability of being in class c, and π_{ic} is the probability of a correct response for item i given the student is in class c. Different CDMs provide different parameterizations for calculating π_{ic} (Rupp, Templin, & Henson, 2010).

CDMs can differ in several different ways. Some of these include the type of observable variables that can be modeled, the type of latent variable model that can be used and/or how different skills can be combined. The type of variables are generally either dichotomous or polytomous. The relationship between different attributes, with regards to how each influences the probability associated with a particular observable variable that depends on multiple attributes, can be modeled in either a compensatory (having one attribute makes up for having a lack in the second attribute) or a non-compensatory manner as to how a given observable variable depends on their values. Some models may be more appropriate for certain designs. Von Davier (2008) expanded this idea into a general diagnostic model. This model follows the same principle as the CDMs described above regarding the fact that there is a mapping (the Q-matrix) of attributes to items. The difference is that this general model allows for polytomous item response as well as polytomous attribute variables. The general formula for this diagnostic model (following the parameterization specified above) is:

$$P_i(x \mid a) = P(x \mid \beta_{ig}, q_i, \gamma_{ig}, a) = \frac{\exp(\beta_{xig} + \gamma_{xig}^T (h(q_i, a)))}{1 + \sum_{y=1}^{m_i} \exp(\beta_{xig} + \gamma_{xig}^T (h(q_i, a)))} \quad \text{(Von Davier, 2008)}$$

Where i indicates the item in question, g is the population, x is the response pattern, a is the attribute pattern, and β is the difficulty value of each item (which could vary across different populations). The term $\gamma_{xig}^T (h(q_i, a))$ represents how the probability changes as a function of the attributes (a) the subject has. For this formula, γ is a weight vector (transposed) and the $h(q_i, a)$ indicates how much of the attribute the subject has based on the Q matrix. The observable variable is a categorical variable. This function can differ depending on different factors, such as the nature of the attribute variables and whether the model is compensatory or non-compensatory (Von Davier, 2008).

Many different CDMs have been developed and several papers have been written exploring some of the difference between these models (Rupp & Templin 2008, DiBello, Rousos & Stout, 2007). However, similarly to the MIRT model the CDMs also are aimed at measuring multiple attributes. While these attributes may provide information about the level of the learning progression, it does not provide direct information on the learning progression. Extensions of CDMs that do reflect learning paths are being developed under the appellation of the attribute hierarchy approach (Leighton, Gierl, & Hunka, 2004), which incorporates structures among student-model variables defined by attribute patterns, which can in turn be interpreted as levels in learning progressions.

Bayesian Inference Networks

One type of model which has been considered a CDM is a Bayesian Inference Network (BIN). BINs are different from other CDMs in that they are a framework versus a specific model. Because of that, BINs are more flexible than other cognitive diagnostic models. However, with the choice of using a BIN comes more decisions regarding how the network is modeled.

A BIN is a graphical representation of the relationships between variables. It is based on a finite acyclic directed graph (Almond, Dibello, Moulder, & Zapata-Rivera, 2007). In general, a graph is a set of vertices (V) and edges (E), where an edge is a line between two vertices. An edge can be represented by the two vertices it connects such as (V1, V2). A finite graph is one with a finite number of vertices. A directed graph is one in which the edges are directed i.e. the edge (V1, V2) is different from the edge (V2, V1) as these two edges would indicate a different type of dependency. In graph theory the arrows imply direction, as in if the line (V1, V2) is included but not (V2, V1) then this would mean that starting at vertex V1 movement is allowed to V2. However, starting at V2 movement is not allowed to V1 using the edge connecting the two vertices. A path is a set of edges in which the starting vertex for an edge is the same as the ending vertex from the previous edge. An acyclic graph is one in which there is no path that goes from one vertex back to itself.

One example of a BIN with 5 variables in which variables 3, 4 and 5 are dependent on variables 1 and 2 can be seen in Figure 6.

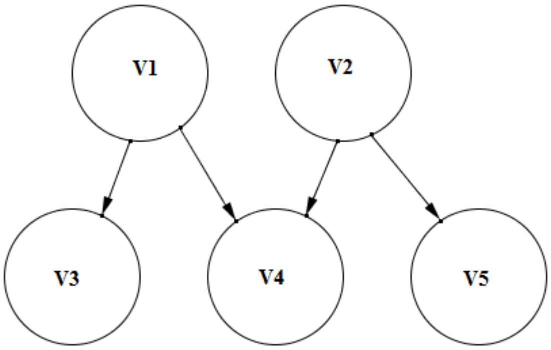


Figure 6: A basic BIN with 5 variables, variables 3, 4 and 5 are dependent on variables 1 and 2

In a BIN the vertices are thought of as categorical variables with values representing states. A given person is thought to be in one state, represented by one possible value of the categorical variable. The dependency represented by an edge is a probabilistic dependency, so the edge (V1, V3) (as seen in Figure 6) would imply that the probabilities associated with the states in V3 differ depending on the state of V1. Or put another way, the probability of V3 is conditionally dependent on V1. For the edge (V1, V3) V1 is referred to as the parent node, and V3 is called the child node.

Nodes in a BIN may have no parents, one parent or multiple parent nodes. The probability distribution associated with each node is conditionally dependent on all of its parents nodes:

$$P(X_i = x_i) = P(X_i = x_i \mid pa(X_i))$$

For a given set of response states the joint probability is :

$$P(X_i = x_1, ..., X_n = x_n) = \prod_n P(X_i = x_i \mid pa(X_i)) \quad \text{(Almond et al., 2007)}$$

Here $pa(X_i)$ represents the parents of node X_i. A BIN is considered to be built when all of the probability distributions for the variables have been determined. The joint product of the conditional probabilities of all variables given their parents (interpreted to include marginal distributions for variables that have no parents) is a joint probability distribution for the full set of variables. At this point a person may enter any information they know and the probabilities will be updated (as shown in the examples below) to determine the probability of each of the unknown variables taking on different values.

In a very simple example, a BIN can be constructed to represent the relationship between the weather and whether or not I take an umbrella with me to work. For this example there are two variables. Variable A is the weather and for this example it can take on the values of sunny, rainy, cloudy, and snowy. The other variable is the variable for if I take an umbrella with me and it can take on the values yes or no. The graph for this is represented in Figure 7. Notice in the graph that the umbrella variable is dependent on the weather variable (made clear by the arrow pointing from the weather variable to the umbrella variable). This arrow indicates that whether or not I take an umbrella is dependent on the weather. It would be a very different statement if the arrow pointed the other way. Using that direction, the BIN would indicate that whether or not I take an umbrella has some influence on the weather.

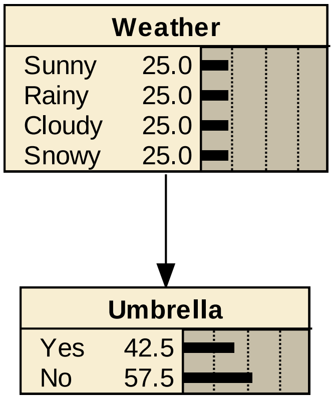


Figure 7: BIN for the relationship between two variables. In this case the probability of an umbrella is dependent on the weather. Shown is the starting conditions when neither value is known.

Each variable has its own probability table. For the weather variable this is the probability of each type of weather occurring (see Table 4). For the umbrella variable this is the conditional probability given the type of weather (see Table 5). While this data is hypothetical, in general these probabilities would come from theory or they would be derived from real data.

Table 4: Probability of a given type of weather

Weather	Prob
Sunny	25%
Rainy	25%
Cloudy	25%
Snowy	25%

Table 5: Conditional probability of taking an umbrella given the type of weather.

	Umbrella	
Weather	yes	no
sunny	10%	90%
rainy	90%	10%
cloudy	50%	50%
snowy	20%	80%

In the initial state the type of weather is not known and whether or not I took an umbrella is also not known. The probability for the weather variable is simply the starting probability for this variable (which could be based on knowing the season, a current weather forecast, or simply looking out the window). The probability that I took an umbrella is the marginal probability across the possible weather conditions and is calculated by:

$$P(u = x_i) = \sum_{W} p(u = x_i \mid w_j) * p(w_j)$$

Where u is the umbrella variable, x_i is either yes or no, W is the weather variable and w_j is either sunny, rainy, cloudy, or snowy. Using Table 4 and Table 5 the probability of a yes is equal to: (.10)(.25) + (.90)(.25) + (.50)(.25) + (.20)(.25) = .425. And similarly the probability of a no is equal to (.90)(.25) + (.10)(.25) + (.50)(.25) + (.80)(.25) = .575.

Once the value of the weather variable is known then the umbrella variable can be updated by using the conditional probability table. If for instance it is raining then the probability that I took an umbrella becomes a .90 and the probability that I did not take an umbrella is .10 (see Figure 8). In general, once the value of a parent node is known the probabilities of the child node follow the conditional probability table for that value of the parent node. Updating can also be done in reverse, if the child node is known then this can modify the probability of the parent node. This type of updating will be discussed in the next example.

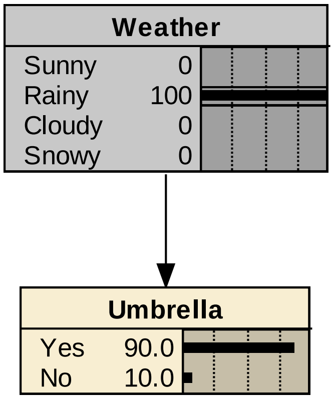


Figure 8: BIN for the relationship between two variables when one is known. In this case it is known that it is rainy, which then implies that the probability of taking an umbrella is 90 %.

In an educational setting a BIN may be constructed for an assessment. In the simple case there is one attribute that is being measured, and each of the items on the exam are designed to measure an aspect of that attribute. A traditional assumption in item response theory (IRT) is that items are locally independent, meaning that the responses to any two items are independent given the student's ability. This same assumption can be made in a BIN by having each of the items depend on the attribute without any direct dependencies among them (see Figure 9).

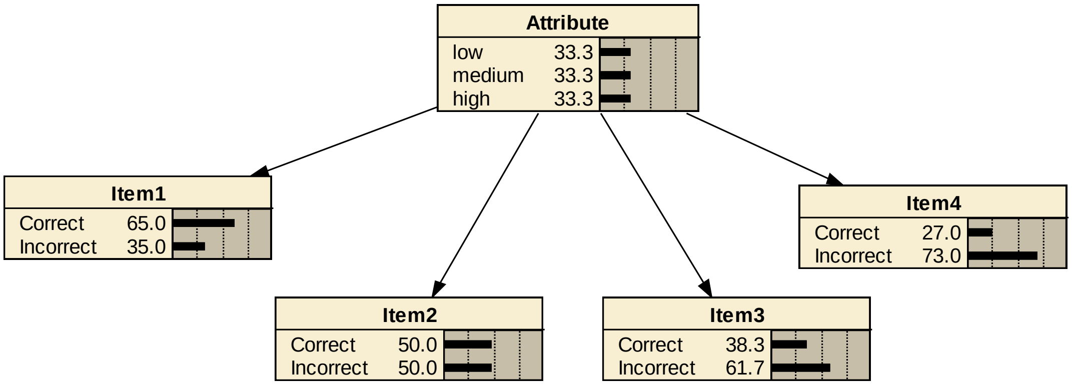


Figure 9: BIN for an IRT model with four items depending on one attribute

The probability of responses (for this example either correct or incorrect) depends on the attribute level of the student. While in IRT this attribute is represented as a continuous latent variable, for a BIN it should be categorical. Different situations may call for different methods of categorizing this variable. For this example, the attribute has been modeled as being able to take the values low, medium, and high. The initial probability of a student being at any of these levels is the same across levels. Items may have different probability structure from each other. In this example we have four different items with different conditional probabilities (see Table 6).

Table 6: Conditional probabilities of correct responses given attribute level

Attribute	Item 1		Item 2		Item 3		Item 4	
	correct	incorrect	correct	incorrect	correct	incorrect	correct	incorrect
low	25%	75%	20%	80%	10%	90%	1%	99%
medium	80%	20%	40%	60%	20%	80%	20%	80%
high	90%	10%	90%	10%	85%	15%	60%	40%

Notice that the overall probability of a correct response for the items decreases as the item number increases. This can be seen in the conditional probabilities of the items as well. For Item 1, most people at a medium or high level should get the item correct. For Item 2 a student can get the item correct if they are at a medium level but are still more likely to obtain an incorrect answer. However, at a high level a student should be getting the item correct. This indicates that the level required by Item 1 is only medium while Item 2 requires a high level of understanding. Item 3 is similar to Item 2 in the level of attributes that it requires, but has slightly lower probabilities indicating that it may be more difficult than Item 2. Item 4 is even more difficult still as even at a high level, the chance of getting this item correct is only slightly over 50%.

The overall probabilities of obtaining an item correct (see Figure 9) were found as noted in the previous example. If the response to item 1 is now known to be correct then this would modify the overall probabilities as seen in Figure 10. Notice that knowing they got a correct response to Item 1 reduced the probability that the student is a low level and increased the probability that they are at a medium or high level. This type of updating is performed using Baye's rule.

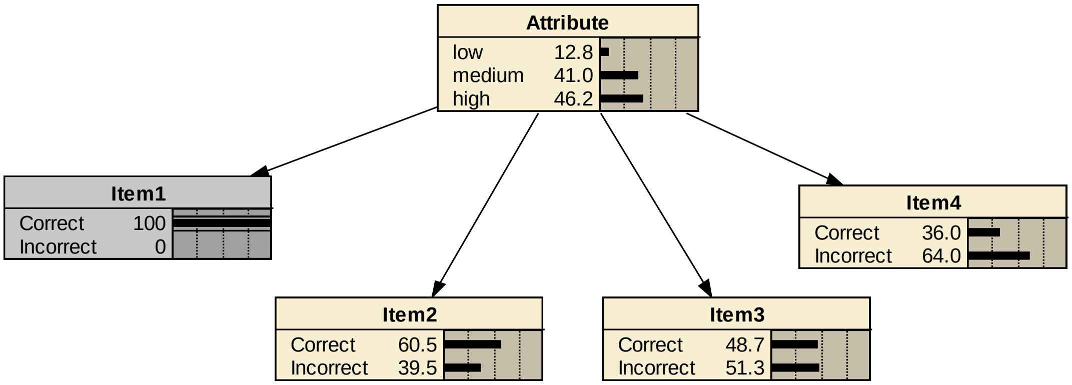


Figure 10: BIN for an IRT model with 4 items, answer to one item is known

Baye's rule states that for any two events A and C:

$$p(A|C)=\frac{p(C\,|\,A)p(A)}{P(C)}$$ (Koski & Noble, 2009)

For this example this can be written as

$$p(X_i\,|\,Y_j)=\frac{p(Y_j\,|\,X_i)P(X_i)}{\sum_X p(X_i)p(Y_j\,|\,X_i)}$$ where X_i is the skill level of the student and Y_j is the outcome of the item in question.

This is often written as:

$p(X_i\,|\,Y_j)\propto p(Y_j\,|\,X_i)P(X_i)$ which can be stated as the posterior distribution (the updated probabilities of the attribute level) is proportional to the likelihood (how likely is the outcome that has been received given the prior probabilities of the attribute level) times the prior distribution (the previous belief regarding the probabilities of each attribute level).

For our example, once we know that a correct response was found, then the likelihood of the attribute level becomes the values from the table that correspond to a correct response. This can then be multiplied by the initial probabilities and then normalized (by dividing by the total of this column) to obtain values that sum up to 1. The result gives the posterior probabilities (see Table 7). Notice that these are the probabilities indicated as the probabilities for the skill level in Figure 10.

Table 7: Updating the attribute level probability based on a correct response to item 1

Attribute	Conditional probability of having item 1 correct	likelihood	Prior Prob.	likelihood *prior	Norm. coeff.	posterior (likelihood * prior/ normalizing coefficient)
low	0.25	0.25	0.33	0.0825	0.6435	0.128
medium	0.8	0.8	0.33	0.264	0.6435	0.41
high	0.9	0.9	0.33	0.297	0.6435	0.462

This updating can be done as the response to each item is found. In Figure 11 we see the resulting probabilities for a student who obtained correct responses to Items 1 and 2 and incorrect responses to Items 3 and 4. From this table, if we had to indicate a single category to categorize this student we would say that student has a medium level of ability as that is the category with the highest probability.

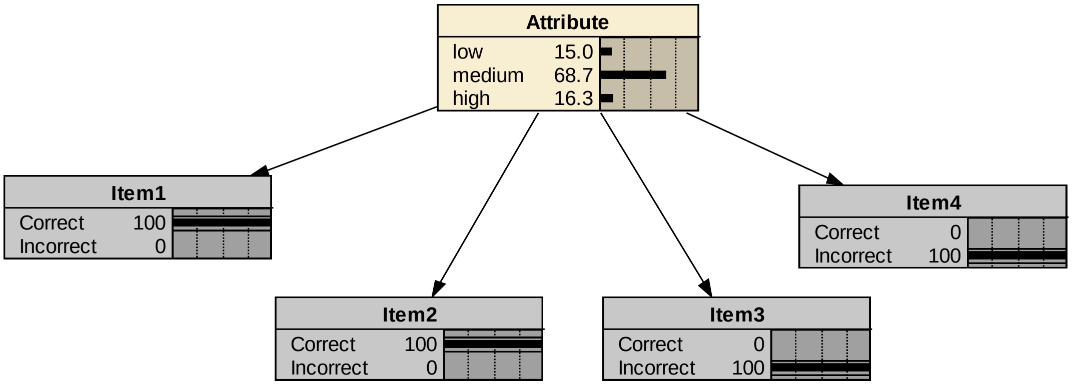


Figure 11: BIN for an IRT model with four items with known results.

Notice also that when the answer to only one item was known the probabilities were updated for the other items. This is due to the fact that the probability for the attribute level was modified which then modified the overall probabilities of each item (using the straightforward method seen in the previous example.)

This process is slightly more complicated when there are multiple parents but the general concept is the same. One of the uses of a BIN is that the probabilities are updated even if only partial information (such as knowing the student's responses to only some of the questions) is known. Due to this fact, a student's state can be estimated even if only partial information is known. Examining the probabilities may also provide information on the strength of the belief in this estimate. For example, if estimates for two different groups are fairly similar than one may not want to conclude that a student is in one group over the other even if the probability is slightly higher for one. However, if the probability for a student being in one group is fairly high and for the others it is fairly low, then one would have more confidence in the categorization of that student.

Following Liu (2009), this example can be expanded to the case of learning progressions where the parent node is the learning progression or the attribute that is being measured and the child nodes are the item nodes (assuming again that there is one test with several items that measure the learning progression). Again here the items are independent given the attribute level of the student. This network would look the same as the one in the previous example, the difference being in the interpretation of the learning progression (as opposed to just one attribute). Another way to represent a learning progression may be to break it up into different attributes and have each item depend on the attributes needed to complete this item. This leads to further questions regarding how to model the relationship between attributes. These different techniques will be discussed in Chapter 3.

Liu (2009) creates GC (group by composite skills) matrices based on competence patterns; which are matrices similar to the Q matrix, except that while the Q matrix lists which attributes each item depends on, the competence patterns lists different type of people by what attributes they have. These attributes could be basic (represented by a single letter) or composite (represented by multiple letters) (see Table 8). The composite attributes indicate that students are able to integrate the basic attributes involved in the composite. The columns represent the basic or composite attribute, while the rows represent different groups of students. These groups are defined based on the possible learning trajectories of students. For example, in Table 8 there are 7 possible groups. The first group represents students who have only mastered the first attribute (A). In Group 2 students have mastered attribute A and B while in Group 3 students have mastered attributes A and C but not B. Group 4 assumes students have mastered the first three attributes, while groups 5-7 display mastery of the final attribute (D). These response patterns arise from three different learning trajectories: one where students acquire attribute D by first acquiring A and then B, a second where they learn D by first acquiring A and then acquiring C, and a final trajectory where they acquire A and then both B and C (in either order) and then D.

Table 8: An example GC matrix

person pattern	A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD
G1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G2	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
G3	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0
G4	1	1	1	0	1	1	0	1	0	0	1	0	0	0	0
G5	1	1	0	1	1	0	1	0	0	0	0	1	0	0	0
G6	1	0	1	1	0	0	0	0	0	0	0	0	1	0	0
G7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

This matrix relates to learning progressions by associating different competence patterns to given levels of a learning progression. While it may be the case that there are multiple competence patterns allowable in the same level of the learning progression, each pattern should only fit at one level. The relationship between levels of the learning progression would depend on the hypothesized competence patterns and how they relate to each other. These patterns should represent how students grow with more complex competence patterns at higher levels of the learning progression. This may be a very simple relationship, as in the case where each attribute is believed to be learned in progression. Therefore the possible competence patterns would be if students have mastered an attribute then they should also have mastered any attributes at a lower level. In this case, the learning progression may have a single level for each of the attributes in the patterns. For the example in Table 8, a three level learning progression may be hypothesized. The first level would correspond to having attribute A, the second to having A and either B or C along with either composite AB or AC. It could also include having both B and C and the composite BC. The final level would add in attribute D as both a basic attribute and as part of the composite attributes.

If everyone always responded correctly to items that the state they were in implied they should be able to answer, and always answered incorrectly to items they did not have the attributes for then the conditional probabilities would be 1 if they had the attribute needed and 0 if not. This, however, is not often the case, as students tend to make errors sometimes and answer incorrectly to items they should be able to answer (which can be referred to as a slip (Liu, 2009)) or they could guess and answer correctly to an item for which they really do not have the underlying abilities (Liu, 2009).

This is similar to the concept of error in latent class scaling analysis. In latent class scaling analysis the underlying latent ability has several different levels. Depending on student's level, there are expected responses to a certain set of items. For example, a questionnaire that is designed to measure a person's attitude regarding the death penalty might have questions that range from "All criminals should be sentenced to death" to "No criminals should be ever be given the death sentence". It may be posited that there are 3 types of people, those who believe the death penalty is never appropriate, those who believe it is appropriate in extreme circumstances, and those who think it should be the sentence more often. For each of these three types of people (which would be interpreted in the model as a latent class with 3 ordered levels) there is an expected response pattern, with people in category 3 being more likely to agree with the more extreme statements and people in category 1 would be less likely to agree with those statements.

When dealing in latent class analysis it is common to put some constraints on the model such as setting the error probabilities to be equal to one another. Similarly, in a BIN, models may be constrained in order to aid in estimation. Equality constraints may be used or particular probabilities may be set. Other constraints may be regarding the probability distribution for a given relationship. One benefit in using BINs is that there is much flexibility in how these constraints could be set and which constraints could be used. Their flexibility makes them very useful for diagnostic modeling (Almond et al, 2007).

In fact, constraints can be made that could incorporate other CDMs into the Bayesian Network framework. When estimating the conditional probabilities, constraints could be placed on the model that would incorporate the parameters associated with a specific CDM. These parameters would be estimated and from their estimation the conditional probabilities can be calculated.

For example, one common CDM is the deterministic input, noisy-and-gate (DINA) model. This is a non-compensatory model, which implies that a student must have all of the attributes required of an item in order to have a high probability of answering that item correctly. The lack of one of the attributes cannot be made up for by having another attribute. In this case:

$\pi_{ic} = (1 - s_i)^{\xi_{ic}} \; g_i^{1-\xi_{ic}}$ where ξ_{ic} is 1 if the student is in a class that has mastered all of the attributes required by the item, and 0 otherwise, s_i is the slipping parameter which is the probability of getting the item incorrect given the student has the correct attributes, and g_i is the guessing parameter, which is the probability of getting the item correct given the student does not have the attributes required (Rupp, Templin, & Henson, 2010). The Q-matrix can be used to find ξ_{ic} by matching the row in the Q matrix that corresponds to the item with a vector that contains the mastery of the attribute for the students. If for all the 1's in the Q-matrix the student also has mastery of that attribute then the result is a 1.

A graphical representation of how this model would look in a BIN can be seen in Figure 11. There are four attributes and four items, each item depending on a different set of attributes. Each attribute can be thought of as a latent variable having two classes, one class indicating the student has the attribute and the other indicating that the student does not have the given attribute. The lines indicate that whether or not the student has the attribute has an effect on the students' probability of a correct response. The guessing and slipping parameters can be estimated, and the conditional probabilities can be calculated using the formula for the DINA model ($\pi_{ic} = (1-s_i)^{\xi_{ic}}\ g_i^{1-\xi_{ic}}$), where ξ_{ic} can be determined by the corresponding combination of attributes.

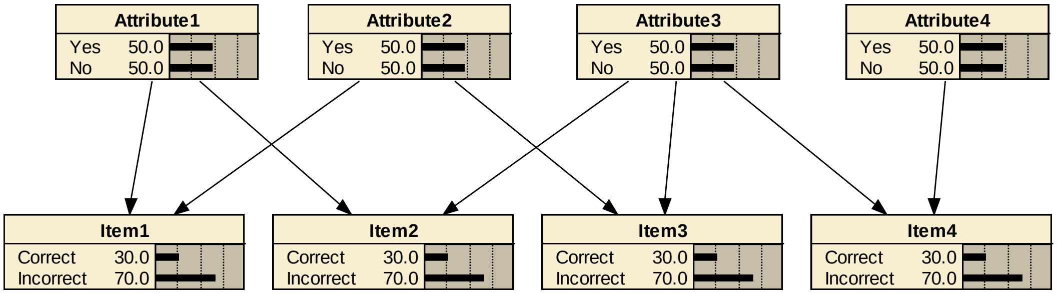


Figure 12: A BIN based on the DINA model

For example if the slipping parameter was .1 and the guessing parameter was .2 then for Item 1 the corresponding conditional probability table can be seen in Table 9. Notice that the probabilities are the same if the student only has either one of the attribute or neither of the attributes as both attributes are required elements of the item.

Table 9: Probability of item responses to item 1 in a DINA model.

Attribute 1	Attribute 2	Item 1	
		Correct	Incorrect
Yes	Yes	0.9	0.1
Yes	No	0.2	0.8
No	Yes	0.2	0.8
No	No	0.2	0.8

The disjunctive version of this model is the deterministic input, noisy-or-gate (DINO) model. While this model is not as popular, it will be used here to demonstrate the differences between a compensatory (of which a disjunctive model is one type) and a non-compensatory model, as in Study 2 these two types of models will be compared. This model also uses the slipping and guessing parameters as described above. The difference is that

$\pi_{ic} = (1-s_i)^{\omega_{ic}} g_i^{1-\omega_{ic}}$ where $\omega_{ic} = 1-\prod_A (1-a_{ca})^{q_{ia}}$ where a_{ca} is 1 if the student is in a class which has the attribute in question and 0 otherwise and q_{ia} is 1 if the item requires that attribute, and 0 otherwise. In other words, ω_{ic} is 1 if the student has at least one of the required attributes and 0 if they have none of the required attributes.

The graphical structure of this model is the same as that for the DINA model (as seen in Figure 13). This model again uses guessing and slipping parameters which can be estimated. The difference is that this estimation will produce different overall conditional probability tables. The result of a conditional probability table where the slipping parameter is again .1 and the guessing parameter is .2 can be seen in

Table 10.

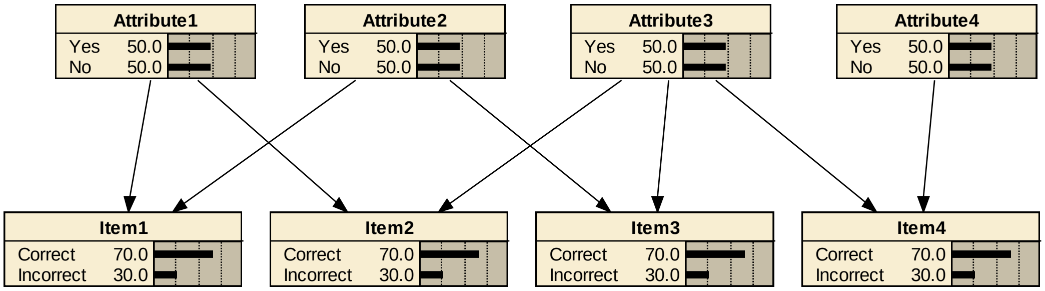


Figure 13: A BIN based on the DINO model

Table 10: The conditional probabilities for item 1 based on the DINO model

Attribute 1	Attribute 2	Item 1	
		Correct	Incorrect
Yes	Yes	0.9	0.1
Yes	No	0.9	0.1
No	Yes	0.9	0.1
No	No	0.2	0.8

While there may be many different choices in the realm of cognitive diagnostic models that can be used to measure learning progressions this research will focus on BINs. Chapter 3 will provide specific information regarding how this type of model can be applied for the purpose of modeling learning progressions and how different Bayesian networks may be compared.

MCMC Estimation

In recent years there has been an increase in the use of Markov Chain Monte Carlo (MCMC) estimation, in particular for complex models (Sinharay, 2004). The mechanics of MCMC estimation make it particularly suitable to Bayesian estimation problems, although it can be applied to other non-Bayesian applications such as likelihood analysis or decision theory as well (Mignami & Rosa, 2001). With the increase in sophistication of computer programs this technique can be fairly straightforward to implement as there are computer programs that have the MCMC algorithm already programmed into them (such as Winbugs and R).

Recall from previous in this chapter the application of Baye's rule in a BIN:

$$p(X_i \mid Y_j) = \frac{p(Y_j \mid X_i)P(X_i)}{\sum_X p(X_i)\,p(Y_j \mid X_i)}$$

This equation holds when the range of possibilities for X is finite. However, if the values that X can take are continuous then the equation becomes:

$$p(X_i \mid Y_j) = \frac{p(Y_j \mid X_i)P(X_i)}{\int_X p(X_i)\,p(Y_j \mid X_i)dX}$$

If we examine the overall probability where D is the data that is observed, and θ is the distribution of values of the parameters then this would be represented as:

$$p(\Theta \mid D) = \frac{p(D \mid \Theta)P(\Theta)}{\int p(\Theta)\,p(D \mid \Theta)d\Theta}$$

When estimating values, a particular function of the distribution is of interest (such as examining the mean of the distribution, to obtain what are called “expected a posteriori” or EAP estimates of the parameters). This can be represented as f(θ) and the posterior expectation of this function is given by:

$$E[f(\Theta) \mid D] = \frac{\int f(\Theta)\,p(\Theta)\,p(D \mid \Theta)d\Theta}{\int p(\Theta)\,p(D \mid \Theta)d\Theta}$$ (Gilks, Richardson, & Spiegelhalter, 1996).

This expectation can be quite difficult to compute. Monte Carlo integration avoids this difficulty by producing a discrete approximation of the expected values. It does this by drawing samples from the distribution and taking the average of these samples. If the samples are independent then as the number of samples increase the approximation becomes more accurate (Gilks, Richardson & Spiegelhalter, 1996).

In order to draw samples a Markov chain is used (hence the term Markov Chain Monte Carlo). The definition of a Markov chain is a sequence of random variables ($\Theta^1, \Theta^2, ...\Theta^t$) such that the distribution of Θ^t given all previous Θ depends only on the most recent value, Θ^{t-1} (Gelman et al., 2004). Each chain starts at an initial value and then each Θ^t is drawn from a transitional distribution $T_t(\Theta^t \mid \Theta^{t-1})$. This transitional distribution is the conditional distribution for the parameter in focus, given the data and treating the previous draws of all other unknown parameters as true. The concept is that each draw gets the accumulated distribution of all draws thus far thereby getting closer to the distribution of interest, and after a sufficient number of draws (or burn-in) the probability distribution for a draw from the chain will converge to the probability distribution of interest, and the accumulated distribution of draws converges to that distribution (Gilks, Richardson, & Spiegelhalter, 1996).

This study will use the Winbugs computer program (Spiegelhalter, Thomas, Best, & Lunn, 2003) to perform the MCMC estimation. While there are other programs that could be used to estimate BINs (such as Netica (Norsys Software Corporation, 2007) and Genie (Decision Systems Laboratory, University of Pittsburgh, 2003)), Winbugs is a flexible program that will allow for the use of constraints when estimating the conditional probabilities.

Winbugs uses the Gibbs sampling algorithm and various univariate samplers within Gibbs to generate the sample draws for the Markov chain. This algorithm has been proven to converge to the distribution of interest under broadly satisfied conditions (Gilks, Richardson, & Spiegelhalter, 1996). For each time point, the algorithm samples from the transitional distribution and the new sample is considered to be a candidate point. This candidate point is accepted or not accepted and if not accepted than the previous sample is used and the chain does not move (Gilks, Richardson, & Spiegelhalter, 1996).

CHAPTER 3: MODELING LEARNING PROGRESSIONS WITH BAYESIAN INFERENCE NETWORKS

While the true underlying model of an assessment (i.e. the structure and nature of the relationships between the observable variables and the latent variables) may not be known, there are many different choices that can be used to try to approximate its true nature. The theory behind the development of the learning progression can help guide the decisions regarding the structure of the model. In addition, research can provide insights into the implications for different model choices by examining different models under known circumstances. This chapter will discuss different relationships and modeling techniques.

Setting up the Bayesian Inference Network

As discussed in the previous chapter, a learning progression may follow from a learning path. In the case of a learning path, where the levels are in a progression (as in Figure 4), a simple Bayesian model may have one latent variable representing the learning progression associated with the observable variables (see Figure 14). Note that this model is a discrete approximation of a continuous unidimensional IRT model, where probability restrictions have been imposed to approximate the ability continuum as a categorical variable with four values.

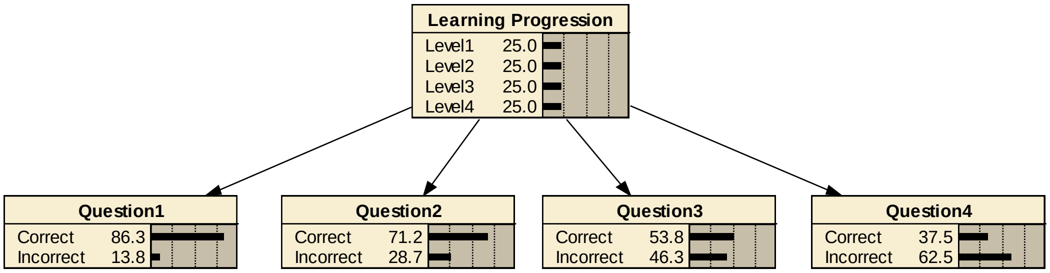


Figure 14: An example BIN with one variable representing the LP

The latent variable has different categories each associated with a different level of the learning progression. A hierarchical structure may be imposed in this model by making the probability of obtaining a correct response higher for students in higher classes (see Table 11 for an example conditional probability table). For this type of model, it is assumed that each student belongs to exactly one latent class and probabilities for a correct response depends on that class.

Table 11: An example of the probabilities for an item that depends on an LP. (Note how the probabilities increase as the level of the student increases)

Question 3	Learning progression			
	Level 1	Level 2	Level 3	Level 4
Correct	20%	30%	80%	90%
Incorrect	80%	70%	20%	10%

Another way of modeling the learning progression is to treat each level of the learning progression as independent (see Figure 15). This may be more appropriate when the learning progression consists of different attributes, each associated with different levels of the learning progression, and which students may learn in varying orders. In fact, modeling the attribute variables as independent posits that no path is any more likely than any other path. (This assumption is usually not tenable, and more constrained versions with probabilistic associations will be discussed later.) Items are then targeted at measuring specific levels of the learning progression. This model would be equivalent to a discrete MIRT model with four dimensions (each dimension representing a level of the learning progression) and a simple structure.

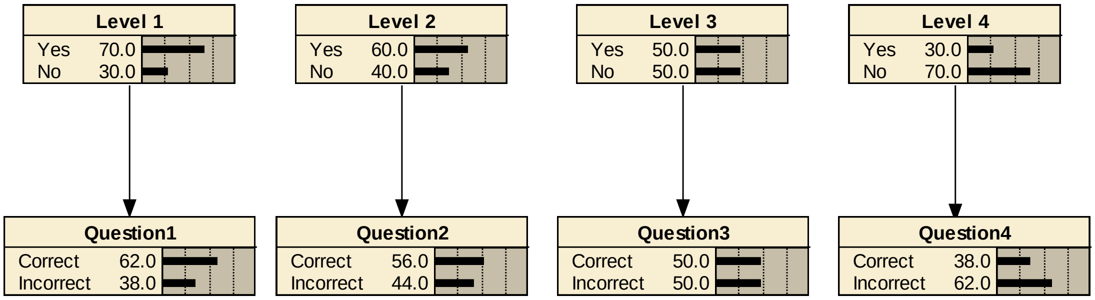


Figure 15: An example BIN with separate variables representing each level of the LP

One place where this type of model may be more appropriate is in a facet based approach to a learning progression. In this type of approach different facet clusters are identified, each with a goal level of understanding and including some problem levels of understanding (such as misconceptions) (DeBarger et al., 2009). While there may be some belief regarding how students learn the given facets, there is not necessarily a strict progression.

One example of a facet cluster for the model of an atom is shown in Table 12. Note that there is one main facet which consists of three goals (or attributes that are desirable in the student). There are also four problems that the students may have. Note that the problems can be associated with a lack of one (or more) of the goals. (Problem 3 would correspond to lacking parts of both goal 1 and goal 2).

Table 12: An example of a facet approach LP for the model of an atom (taken from DeBarger et al, 2009)

Facet	The student correctly uses a model for the atom to account for the structure of matter
Goal 1	The student knows most of the mass of the atom is in the nucleus, which is made up of protons and neutrons
Goal 2	The student knows that electrons move outside of the nucleus and that the space the electrons move in defines the volume of the atom
Goal 3	The student understands that atoms are electrically neutral when they have an equal number of protons and electrons
Problem 1	The student has an incorrect model for the charge of parts of the atom
Problem 2	The student has an incorrect model for the mass of the parts of an atom
Problem 3	The student has an incorrect model for the location of parts of the atom
Problem 4	The student has an incorrect model for the size of the parts of the atom

A learning progression could be generated from this model in the sense that at the highest level the student has all of the goal facets, while at the lower level the student is missing one or more of the goals (See Table 13 for an example). However, the learning progression would then assume that students learn about atoms in a particular order, and in order to move into the next level in the learning progression, they go from an incorrect model to a correct model for one of the attributes. This might not be the case, and in fact students may learn in different stages. A more appropriate learning progression might then be one as seen in Table 14. These two learning progressions are similar, with the main difference being that the levels of the learning progression in Table 14 do not state that the student has the attributes from the previous levels. For this type of learning progression it might make more sense to model the individual pieces (as in Figure 15).

Table 13: An LP for the model of an atom with a hierarchical structure

LP Level	Student's ability to model an atom
1	The student knows the location of the parts of an atom
2	The student has the level 1 attribute and the student knows most of the mass of the atom is in the nucleus, which is made up of protons and neutrons
3	The student has the level 2 attributes and the student knows that electrons move outside of the nucleus and that the space the electrons move in defines the volume of the atom
4	The student has the level 3 attributes and the student understands that atoms are electrically neutral when they have an equal number of protons and electrons

Table 14: An LP for the model of an atom with no hierarchical structure imposed

LP Level	Student's ability to model an atom
1	The student knows the location of the parts of an atom
2	The student knows most of the mass of the atom is in the nucleus, which is made up of protons and neutrons
3	The student knows that electrons move outside of the nucleus and that the space the electrons move in defines the volume of the atom
4	The student understands that atoms are electrically neutral when they have an equal number of protons and electrons

Having this type of structure (without a hierarchy) would change the probabilistic dependencies in the model. In this case, the probability for a correct response would only depend on if the student has the attribute for the associated level (as seen in Table 15). Having the attributes for Level 3 does not influence the probability of a correct response to an item that corresponds to Level 2. This more closely follows the learning path shown in Figure 5, as there may not necessarily be a hierarchy in the learning progression. Instead it may be possible that other patterns emerge, such as students having attributes that would correspond to Level 1 and Level 3 in the learning progression, but lacking the attributes required for Level 2.

Table 15: A conditional probability table for an item that depends on one level of an LP

Question 2	Level 2	
	Yes	No
Correct	80%	20%
Incorrect	20%	80%

To state this another way, if four levels of a learning progression could be represented as four different states, such that (0,0,0,0) represents not having any of the skills required for the given levels while (1,1,1,1) would represent a student having all of the skills required. In the hierarchical case with just one latent variable representing the levels then the possible states are: Level 1 would correspond to (0,0,0,0), Level 2 to (1,0,0,0), Level 3 to (1,1,0,0), Level 4 to (1,1,1,0) and Level 5 to (1,1,1,1). In the case where the levels are treated as separate variables this would allow for other states such as (1,1,0,1) which may not have a high probability of occurring but is still possible.

Another instance in which separating out the learning progressions into separate variables may be appropriate is to address the “messy middle” issue. This issue deals with the fact that students do necessarily learn in one trajectory and may in fact display attributes related to a high level of a learning progression but not display attributes of a lower level (Gotwals & Songer, 2009). While it may be straightforward to determine who the novices are, and who the students are that have mastered all levels, it is not always straightforward to determine where a student is on a continuum (Gotwals & Songer, 2009). In this case having information about attributes they display may provide more accurate information about where the student is along the learning progression.

In our addition example from Chapter 2, a student may be able to recognize addition problems and carry when they need to, but they may not be able to add numbers. Examining the overall probabilities for each attribute would determine for which levels the student displays mastery, which could then be used to determine the overall level of the student.

This model may not be very realistic as there generally should be some dependence between the different levels. Another way to model the relationships between the levels is to add dependencies from the lower level to the higher level (see Figure 16).

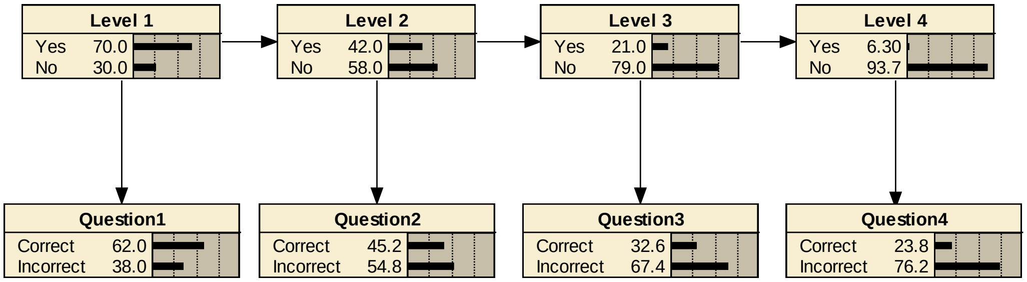


Figure 16: An example BIN with a dependency between the levels of the LP

This dependence could be very strict, where if the student is not at Level 1 (or does not have the Level 1 attributes) then they cannot have the Level 2 attributes. In this case, the conditional probability of having the Level 2 attributes given the student does not have the Level 1 skill set is 0 (see Table 16). Or this dependence could be less strict and these conditional probabilities could be freely estimated. When the dependence is strict this follows the learning path from Figure 4, as how levels are obtained is strictly ordered, whereas if the dependency is lifted then this allows room for the learning path from Figure 5, as it may be possible to have Level 3 skills without having Level 2 skills.

Table 16: Conditional probability of Level 2 depending on Level 1

	Level 2	
Level 1	Yes	No
Yes	60%	40%
No	0%	100%

Whereas the model described above allows the probability of having the attributes for a given level depend solely on whether or not the student has the attributes for the previous level, a model could also be made that would have the attributes depend on all of the previous attributes (see Figure 17).

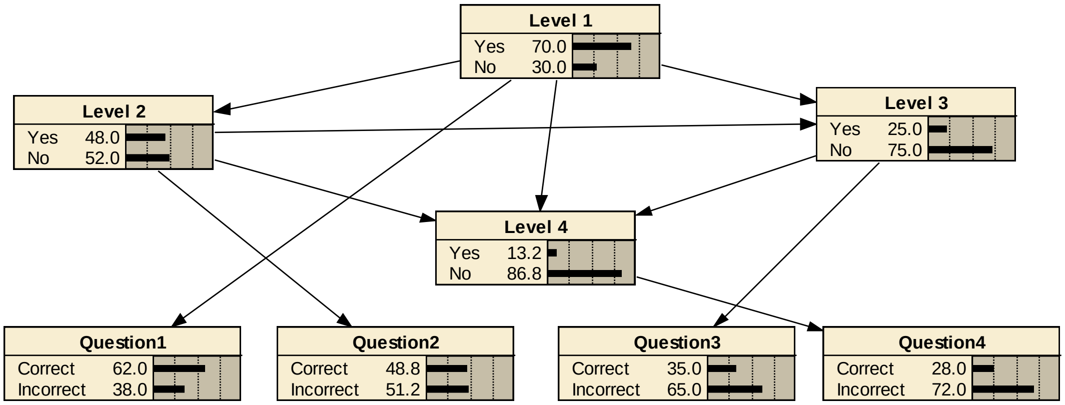


Figure 17: An example BIN with dependencies between each level of the LP

While this may not add much to the previous model, as in either case any combination of levels is possible, it does allow for the case where the absence of a low level attribute might have more effect on higher levels than would just be found by looking at the level immediately below (see Table 17).

Table 17: The conditional probability for Level 3 which depends on Levels 1 and 2

		Level 3	
Level2	Level1	Yes	No
Yes	Yes	50%	50%
Yes	No	20%	80%
No	Yes	10%	90%
No	No	0%	100%

While these models may not be the only models that can be used to measure learning progressions, they reflect different theories regarding the relationship between attributes in a learning progression based on different possible learning paths students can take. Study 1 will be a comparison between the four models presented above that will examine how well each model can recover parameters and classifications, as well as the consequences of misspecification of the model. The details of this study are described in Chapter 4.

Modeling Conditional Probabilities in a Bayesian Framework

For the first model described above, the probability that is to be estimated is $p(j \mid k)$ where j is the response (for a binary observable variable this would be a 0 or 1, for a polytomous observable variable this may take on more values) given that the person is at level k of the learning progression. This conditional probability can be estimated directly or constraints can be placed. An unconstrained model is most flexible, but the number of conditional probabilities to estimate can become excessive and unstable in large problems (Mislevy et al., 2002). One type of constraint with fewer parameters to estimate is to have the probability structure follow an IRT model. This gives the additional benefit of putting the parameters on a familiar scale to experts in educational measurement (Mislevy et al, 2002).

Almond & Mislevy (1999) describe how a BIN can be used to represent an IRT model. While the graphical representation of an IRT model would look on the outset the same as the simple BIN described earlier (see Figure 14), the model used for estimation would include additional item parameters, although a smaller number of parameters would need to be estimated, as these item parameters would be used to determine the conditional probabilities (Almond & Mislevy, 1999). The key difference between the BIN representation and an IRT model is that the BIN has only a finite set of ability parameters values, and is thus a structured latent class model (or in our case corresponding to levels of a strictly ordered learning progression).

The model they follow is the same as the use of the latent class Rasch model (LC/RM) in latent class analysis. Formann and Kohlmann (1998) specify this model as:

$$P\ (X_{ij} = 1) = \frac{\exp(\xi_j + \sigma_i)}{1 + \exp(\xi_j + \sigma_i)}$$

where i indexes the item, and j represents class j (which in our case is the level of the of the learning progression.) The parameters to be estimated are then an item difficulty parameter (σ) and a class parameter (ξ) which is the ability associated with the given levels of the learning progression.

For the case where the item is polytomous this must be expanded in order to include the different possible levels of the item. This can be done using the Samejima - Dibello (Mislevy et al, 2002) model. This model follows the Samejima graded response models, but instead of the person ability being a continuous variable it categorizes the ability into several levels (Mislevy et al., 2002). (Again, here that would correspond to the different levels of the learning progression).

$$P\ (X_{ij} \geq k) = \frac{\exp(\xi_j + \sigma_{ik})}{1 + \exp(\xi_j + \sigma_{ik})} \text{ and } P(X_{ij} = k) = P(X_{ij} \geq k) - P(X_{ij} \geq k+1)$$

In a complex assessment there may be multiple skills required to complete a given task. These skills could each have their own learning progressions. For example, in networking, students may need to be able to perform binary addition as well as configure a router in order to troubleshoot a network activity. Different tasks may require different levels of these two skills.

A very simple Bayesian network for a task that requires two skills is shown in Figure 18. Each of the latent skills (labeled LP for learning progression) will have different stages which represent the different levels of the learning progression. In this case, the probability of a given response on the observable variables (such as Question 1) will depend on the students' level on each of the learning progression.

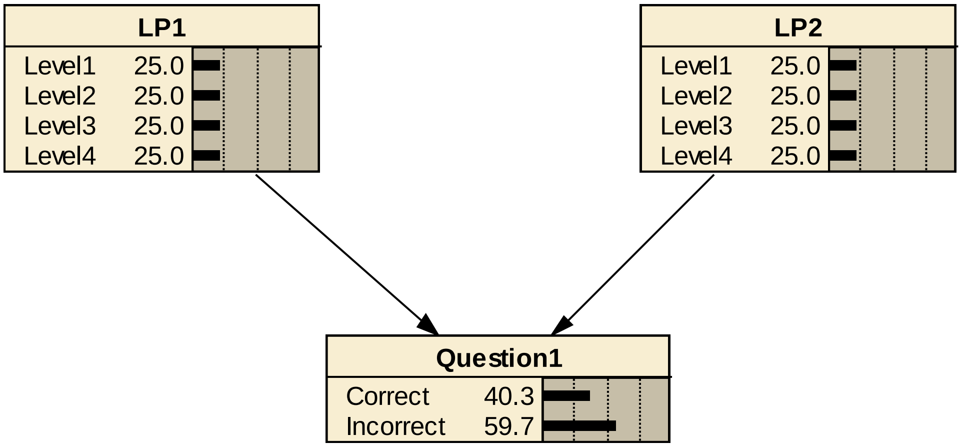


Figure 18: An example BIN with one question depending on two LPs

Here again the probabilities may be estimated directly or constraints can be added to place the parameters on an IRT scale. There are several choices for how the ability of the student on the different skills may be combined to affect the overall probability of a response. Three common types of relationships the abilities can have are compensatory, conjunctive and disjunctive.

In a compensatory relationship the skills complement each other in the sense that having more of one skill makes up for a lack in another skill. Generally the greater ability a person has in each of the skills the greater the probability of a correct response. This is demonstrated in the formula by adding the ability level parameters for each of the skills as follows:

$P\ (X_{i\bar{J}} \geq k) = \frac{\exp(\sum_{J} \xi_j + \sigma_{ik})}{1 + \exp(\sum_{J} \xi_j + \sigma_{ik})}$ where $\bar{J}$ is the vector pertaining to whether or not the student has skill j (for the required skills for the item) (Mislevy et al, 2002).

In a conjunctive relationship the student should have all of the skills required in order to be able to solve the problem. If one of the skills is missing this will hinder the student from solving the problem and having a higher ability in the other skills cannot make up for this lack of skill. In this case the probability of obtaining a correct response is determined by the lowest ability skill as follows:

$$P\ (X_{i\bar{J}} \geq k) = \frac{\exp(\min(\bar{\xi}) + \sigma_{ik})}{1 + \exp(\min(\bar{\xi}) + \sigma_{ik})} \quad \text{(Mislevy et al, 2002)}$$

A disjunctive relationship is one where the highest skill level determines the probability of a correct response. In this sense the ability to solve the problem only depends on the student having one of the skills and does not require all skills. This can be seen as follows:

$$P\ (X_{i\bar{J}} \geq k) = \frac{\exp(\max(\bar{\xi}) + \sigma_{ik})}{1 + \exp(\max(\bar{\xi}) + \sigma_{ik})} \quad \text{(Mislevy et al, 2002)}$$

One note here is that the relationships discussed are conditional on levels of the LPs. This does not address the possible relationships between how students progress through the learning progressions and how the learning of one skill may influence the learning of another skill. The interest is in how the skill levels of the students jointly affect the probability of a correct response at one particular time point. While other types of relationships exist and may be appropriate in different cases, this research will focus on these mentioned here.

There may be some indeterminacy as is typical in IRT analysis. In IRT a shift in the ability measures along with a shift in the item difficulty will produce the same overall probability. Therefore when estimating it is necessary to add a constraint in order to make the estimated values consistent over different runs. There are two types of constraints that are normally used for the Rasch model, one is to center the ability estimates around zero, the other is to center the item difficulty estimates around zero. Following Almond, Yan, and Hemat (2008), this study will center the item difficulty.

While Study 1 will examine the different methods for modeling the structure of the learning progression, Study 2 will examine the different constraints that can be used to structure the relationship between two learning progressions and a set of items which will each be designed to measure both LPs. Study 2 will compare an unconstrained model with the conjunctive, disjunctive and compensatory models for the relationship of the observable variables given two learning progression latent variables, to again examine parameter recovery and misspecification of the models. More details can be found in Chapter 5.

As noted previously, BINs were chosen for this study due to the flexibility they offer in allowing for different models and different constraints. However, other models could have been used instead. For Study 1, an IRT model could be compared to a MIRT model. In the IRT framework cutoff points could be used along the ability framework that would separate students out into levels of the learning progression. This idea could be compared to using the different levels as different attributes in the MIRT framework, each with their own cutoff as to if the student was at that level or not. When multiple LPs are compared then using MIRT could be applied, or the formulas described above for the compensatory, conjunctive and disjunctive models could be applied directly to add constraints into the model.

In a DCM framework, different models could be compared, one that examines a categorical attribute, versus others that separates this attribute out into different binary attributes. In addition, models could be applied (such as the DINA and DINO models) that would put compensatory, conjunctive, or disjunctive constraints onto models measuring two learning progressions.

These studies would be very similar in framework to the current study and may produce similar results. Future studies may want to examine how models compare across frameworks given what model seems most appropriate within a framework. This study chose to focus on BINs in part due to the fact that with a BIN there may be an added benefit in that a probability distribution for where a student is along the LP can be obtained even if the response patterns of the student are not known. Again, follow-up studies may want to determine if this feature would be useful in classroom situations. Having this study would then provide the base for determine the setup of the BIN.

In addition it should be noted that the question asked in Study 1 is different from that asked in Study 2. While Study 1 examines the structure of the variables within a BIN, Study 2 examines the types of constraints that can be placed on a model in a Bayesian framework. The reason why these are different is that having the two studies address different issues should give practitioners a broader view of the type of decisions that can be made when implementing a BIN.

CHAPTER 4: MODELING ONE LEARNING PROGRESSION (STUDY 1)

The first study of this dissertation focuses on different representations of a single learning progression and the relation of the learning progression to the observable variables. Four different representations will be modeled using a BIN framework. Data will be generated by varying parameter settings for each of these models and then each model will be fit to the data sets. The study will compare model fit and student classification rates.

Study Overview

As mentioned previously there are cases in which a hierarchical learning progression may not be the most appropriate representation, as students may follow different learning paths. The question being addressed here is are there certain situations in which it would be beneficial for the purpose of classifying students, to model the learning path as one hierarchical learning progression as opposed to separating the learning progression into different variables and incorporating these variables into a multivariate model.

The case that is examined is where there are multiple observable variables providing evidence about one learning progression. A comparison was carried out among four models.

Table 18 describes the models and includes the probabilities that need to be estimated for each model (for both the latent and the observable variables). Table 19 shows a graphical representation of the models. Note that Model 2 is a constrained version of Model 3 with the constraints P(LP2=1|LP1=1)=P(LP2=1|LP1=0), P(LP3=1|LP2=1)=P(LP3=1|LP2=0) and P(LP4=1|LP3=1)=P(LP4=1|LP3=0). Similarly Model 3 is a constrained version of Model 4, with P(LP3=1|LP2,LP1=1) = P(LP3=1|LP2,LP1=0) and P(LP4=1|LP3,LP2=1,LP1=1)=P(LP4=1|LP3,LP2=1,LP1=0) = P(LP4=1|LP3,LP2=0,LP1=1)=P(LP4=1|LP3,LP2=0,LP1=0)

Also note that Model 1 can be thought of as a constrained version of Model 3 by adding in the constraints that P(LP2=1|LP1=0)=0, P(LP3=1|LP2=0)=0, and P(LP4=1|LP3=0)=0. These constraints make it so that in Model 3 if a student is at a higher level they must have mastered the lower level skills and the probabilities for each level (as represented in Model 1) would be P(LP=0)=1-P(LP1=1), P(LP=1)=P(LP1=1)-P(LP2=1) (i.e. they are at level 1 but not at level 2), P(LP=2)=P(LP2=1)-P(LP3=1), P(LP=3)=P(LP4=1)-P(LP3=1), and P(LP=4)=P(LP4=1). As for the probabilities of the observable variables, knowing that a student had a certain level attributes (such as level 3) in Model 3 would be the same as knowing that the student was at level 3 or above in Model 1.

Table 18: Four models for modeling the relationship between one LP and OVs

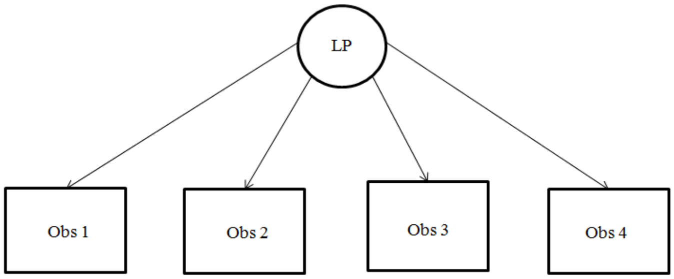
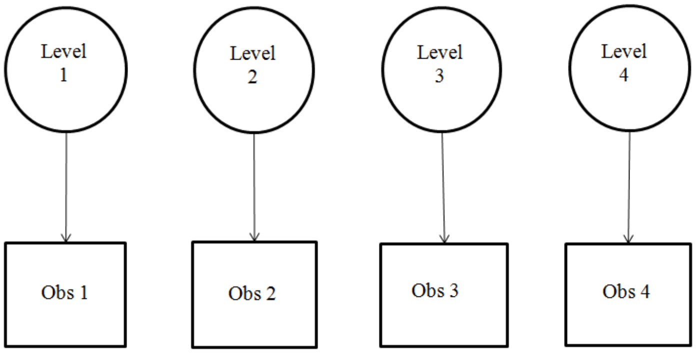
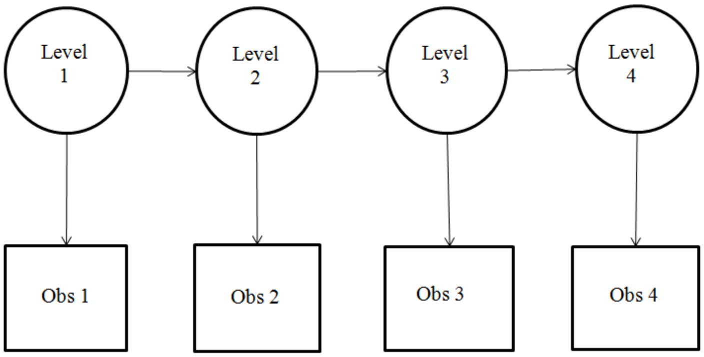
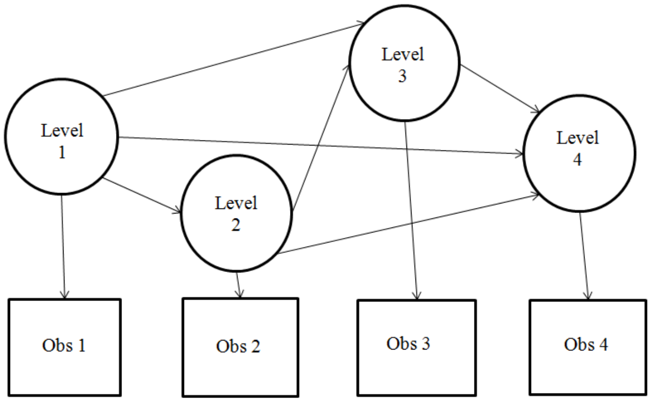
Model	Description	Latent variable probabilities	Observable probabilities
1	One categorical latent variable representing the LP and the observable variables conditionally dependent on that variable	P(LP=i)	P(O(1-12)|LP)

2	4 latent variables representing the individual LP levels, and the observables conditionally dependent on the level they are designed to reflect on. No conditional dependence between the latent variables.	P(LP1=1)
		P(LP2=1)
		P(LP3=1)
		P(LP4=1)
			P(O(1-3)|LP1)
			P(O(4-6)|LP2)
			P(O(7-9)|LP3)
			P(O(10-12)|LP4)

3	Same as model 2 except that each latent level variable is conditionally dependent on the previous, this dependence is freely estimated	P(LP1=1)
		P(LP2=1|LP1)
		P(LP3=1|LP2)
		P(LP4=1|LP3)
			P(O(1-3)|LP1)
			P(O(4-6)|LP2)
			P(O(7-9)|LP3)
			P(O(10-12)|LP4)

4	Same as model 2 except that each latent level variable is conditionally dependent on all previous, this dependence is freely estimated	P(LP1=1)
		P(LP2=1|LP1)
		P(LP3=1|LP1,LP2)
		P(LP4=1|LP1,LP2,LP3)
			P(O(1-3)|LP1)
			P(O(4-6)|LP2)
			P(O(7-9)|LP3)
			P(O(10-12)|LP4)

Table 19: Representative diagram of the different models. Please note that while these diagrams have one observable variable per level the actual simulation will have three observables per level.

Model	Diagram
1	
2	 Level 1 Level 2 Level 3 Level 4 Obs 1 Obs 2 Obs 3 Obs 4
3	 Level 1 Level 2 Level 3 Level 4 Obs 1 Obs 2 Obs 3 Obs 4
4	 Level 1 Level 2 Level 3 Level 4 Obs 1 Obs 2 Obs 3 Obs 4

In order to address the benefit of the models three sub-questions will be addressed:

1) How well are parameters recovered under each model for the various conditions?

2) How do inferences regarding students (i.e., posterior distributions for proficiency variable) compare across the different models under various conditions?

3) How do goodness-of-fit tests perform at identifying the correct model under various conditions?

These questions were addressed by a simulation. Data was simulated based on each of the different models and different parameter specifications using R (R Development Core Team, 2008), and then estimations were computed for each of the parameters using Bayesian inference via MCMC estimation, using Winbugs (Spiegelhalter, Thomas, Best & Lunn, 2003) The resulting parameters were passed back to R for comparison.

In order to estimate the parameters in a Bayesian network, posterior distributions for the variables used in the network are specified. The parameters of interest here are unconstrained conditional probability matrices (or parameters that entail conditional probabilities in lower-order approximations) and distributions of student-model variables. These values for the structural parameters of the BIN model are then used to estimate the probability distributions for student-level variables and observed responses. For each model the probability distributions were specified, the variables (both latent and observable) for a sample of subjects were generated, and then the model parameters were estimated.

Study Conditions

The factors that are varied in the simulation are the sample size and the distribution of the students in the latent classes. The factors that remain the same are the number of levels of the learning progression, the strength of the relationship between the observables and the latent variables, and the number of observables.

The number of levels of the learning progression will be four (with a fifth class representing the novice class for the case when there is just one latent variable). This represents a fairly simple learning progression, while still leaving room for different learning paths. In addition, four seemed to be a common number for the number of levels in a learning progression in the literature (Gunckel, Covitt & Anderson, 2009, Mohan & Anderson, 2009, and Schwarz et al, 2009). Further research may be applied to LPs with more levels.

For each latent variable there are three observable variables that reflect upon the latent variable, making for a total of twelve observable variables. If there are too few observables then the model will not be identifiable (Formann, 2003). In order to keep the model simple but identified, this study followed the approach used by Almond, Mulder, Hemat and Yan (2008) and used three observables per level. Each observable in this study will be dichotomous and will designed to measure a particular level of the learning progression. For Model 1 this means that there will be a jump in the probability of a correct response between students who are at the level below the required level of the item and students who are at the level of LP required by the item or higher. For Models 2-4 this would mean that each observable only has one edge coming into it and that edge is from the level variable that the observable is designed to measure. The relationship between the observable and the latent variable will be set at a medium relationship, which will be represented by using .8 as the probability of a correct response given that the student has the attributes required by the item.

For Model 1 the probability of obtaining a correct response if the student is at a higher level than required is also addressed. There are two conditions, one is that the probability is the same as if the student is at the level required and the other is that the probability increases by .05 (with a max of .95) if the student is at a higher level. These two options reflect the concepts that knowing higher level skills either does (condition 2) or does not (condition 1) aid in solving items designed to measure lower levels of the learning progression.

In addition, the probability of a correct response given the student is below the level required will be .2, which indicates that the student has some probability of answering the item correctly (this is equal to the probability of a multiple choice answer with 5 options). (Other simulation studies have used values between 0 and .3 for this probability (de la Torre, 2009, Liu, Douglas & Henson, 2009) so .2 was deemed an acceptable value.) For Models 2-4 the probability of the observables only depends on whether they have the attributes in question. For this case the probability of .2 will be used for a correct response if they do not have the attributes in question. See Table 20 for the conditional probabilities used for Model 1.

For Model 1, the latent variable parameters that will be used are the probabilities of class membership for each level of the learning progression. In this case there is also a Level 0 that describes students who do not even have the Level 1 attributes.

Table 20: Conditional probabilities for Model 1

conditional probabilities given Level 0

	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob
Cond	1	2	3	4	5	6	7	8	9	10	11	12
1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2

conditional probabilities given Level 1

	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob
Cond	1	2	3	4	5	6	7	8	9	10	11	12
1	0.8	0.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
2	0.8	0.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2

conditional probabilities given Level 2

	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob
Cond	1	2	3	4	5	6	7	8	9	10	11	12
1	0.85	0.85	0.85	0.8	0.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2
2	0.8	0.8	0.8	0.8	0.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2

conditional probabilities given Level 3

	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob
Cond	1	2	3	4	5	6	7	8	9	10	11	12
1	0.9	0.9	0.9	0.85	0.85	0.85	0.8	0.8	0.8	0.2	0.2	0.2
2	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.2	0.2	0.2

conditional probabilities given Level 4

	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob	Ob
Cond	1	2	3	4	5	6	7	8	9	10	11	12
1	0.95	0.95	0.95	0.9	0.9	0.9	0.85	0.85	0.85	0.8	0.8	0.8
2	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8

For this study four conditions will be examined (see Table 21) that represent different distributions of students. The first condition will have equal probability of students being at any level, which may be the case when the exam is administered to a general population. The next condition is one in which the students are mostly high ability, as in the case where most students have studied the material and students who are not prepared would not be taking the exam (such as for a certification exam). A third condition is one where the students are mostly in the middle range of ability. Here the target population is one where students have taken some courses and learned material but they may not have gone the extra step to develop their skills fully, but the exam is also used to determine if there are students who have mastered the higher levels. The last condition is one where students are mostly low ability, such as students who are taking a pre-test for a class and the exam may be used to determine if there are students that will need extra challenges throughout the course.

Table 21: The probability distributions to be used for the LP variable in Model 1

Case #	Description	p(LP0)	p(LP1)	p(LP2)	p(LP3)	p(LP4)
1	Equal probability of any ability student	0.2	0.2	0.2	0.2	0.2
2	Mostly high ability students	0.05	0.1	0.15	0.3	0.4
3	Mostly middle ability students	0.1	0.15	0.3	0.3	0.15
4	Mostly low ability students	0.05	0.4	0.3	0.15	0.1

For Model 2 the probabilities of the latent variable are the individual probabilities of having a given level. The probabilities for Model 1 can be expressed in terms of the probabilities at a given level by accumulating the probability of being in the given class and all of the higher classes (so the probability of having Level 1 attributes is equal to the probability from Model 1 for being at Level 1, 2, 3, 4 or 5, since a student at the higher level should also have the attributes for the lower level). For consistency these probabilities will be used. The case in which students can follow different learning paths is addressed by switching the probabilities for Level 2 and Level 3. Finally, two cases are added, one where there is an equal probability of having Level 2 or Level 3 and one where there is an equal probability of having any of the levels (see Table 22).

Table 22: Probability distributions of having any of the individual level abilities for Model 2

Case #	Description	p(L1)	p(L2)	p(L3)	p(L4)
1	Equal probability of any ability student, following standard progression	0.8	0.6	0.4	0.2
2	Mostly high ability students, following standard progression	0.95	0.85	0.7	0.4
3	Mostly middle ability students, following standard progression	0.9	0.75	0.45	0.15
4	Mostly low ability students, following standard progression	0.95	0.65	0.25	0.1
5	Equal probability of any ability student, reversing levels 2 and 3	0.8	0.4	0.6	0.2
6	Mostly middle ability students, reversing levels 2 and 3	0.9	0.45	0.75	0.15
7	Equal probability of having either level 2 or level 3 skills	0.8	0.6	0.6	0.4
8	Equal probability of having any of the skills	0.6	0.6	0.6	0.6

Model 3 follows similar patterns as Model 2, but now there is the additional condition of students who do not have the previous level attributes. One way this could be modeled is a loose hierarchy, in which case there is a small probability of students having the next level attributes even if they don't have the previous level attributes. For this we would have P(L (X+1)|L(X)=0) > 0 but P(L (X+1)|L(X)=0) < P(L(X+1)|L(X)=1). Another type of condition is to enforce a strict hierarchy by making it so that if a student does not have a level attribute then that student cannot have a higher level attribute i.e. P(L(X+1)|Level(X)=0) = 0. A third method is to make the probability the same of having a level attribute regardless of whether or not the student has the previous attribute i.e. P(L(X+1)|Level(X)=0) = P(L(X+1)|L(X)=1) (which makes this essentially the same as the Model 2).

For this study two conditions will be used when generating data for the condition when the levels are designed to follow the standard progression: that of a loose hierarchy and that of a strict hierarchy. For the case where Level 2 and Level 3 are allowed to switch, a loose hierarchy will be imposed between Levels 1 and 2 but no hierarchy will be imposed between Levels 2 and 3 or 3 and 4 (see Table 23).

For the last model, the same probabilities will be used for the levels if the student has all of the previous levels as they were for Model 3 (see Table 24). For the cases in which a standard progression is expected (Cases 1-4) the following hierarchical structures will be imposed: a strict hierarchy, a loose hierarchy, a hierarchy where Level 1 is loosely required but Levels 2 and 3 are not, and a hierarchy where Level 1 is loosely required and either Level 2 or 3 is required for Level 4 but not both. For the cases in which Levels 2 and 3 are allowed to switch (Cases 5-8) the latter two hierarchical structures will be imposed (see Table 25).

Table 23: Probability distribution for the parameters in Model 3

Case #	Description	p(L1)	p(L2|L1)	p(L3|L2)	p(L4|L3)	p(L2|~L1)	p(L3|~L2)	p(L4|~L3)
1	Equal prob. of any ability, standard progression, strict hierarchy	0.8	0.6	0.4	0.2	0	0	0
2	Equal prob of any ability, standard progression, loose hierarchy	0.8	0.6	0.4	0.2	0.2	0.2	0.2
3	Mostly high ability, standard progression, strict hierarchy	0.95	0.9	0.8	0.5	0	0	0
4	Mostly high ability, standard progression, loose hierarchy	0.95	0.9	0.8	0.5	0.2	0.2	0.2
5	Mostly middle ability, standard progression, strict hierarchy	0.9	0.75	0.45	0.15	0	0	0
6	Mostly middle ability, standard progression, loose hierarchy	0.9	0.75	0.45	0.15	0.2	0.2	0.2
7	Mostly low ability, standard progression, strict hierarchy	0.95	0.65	0.25	0.1	0	0	0
8	Mostly low ability, standard progression, loose hierarchy	0.95	0.65	0.25	0.1	0.2	0.2	0.2
9	Equal probability of any ability, reversing levels 2 and 3, loose hierarchy	0.8	0.4	0.6	0.2	0.2	0.2	0.2
10	Mostly middle ability, reversing levels 2 and 3, minimal hierarchy	0.9	0.45	0.75	0.15	0.2	0.6	0.2
11	Mostly middle ability, reversing levels 2 and 3, loose hierarchy	0.9	0.45	0.75	0.15	0.2	0.2	0.2
12	Equal prob. of having either level 2 or level 3 skills, minimal hierarchy	0.8	0.6	0.6	0.4	0.2	0.6	0.4
13	Equal prob. of having either level 2 or level 3 skills, loose hierarchy	0.8	0.6	0.6	0.4	0.2	0.2	0.2
14	Equal prob. of having either level 2 or level 3 skills, minimal hierarchy	0.8	0.6	0.6	0.4	0.2	0.6	0.4
15	Equal prob of having any of the skills, loose hierarchy	0.6	0.6	0.6	0.6	0.2	0.2	0.2
16	Equal prob of having any of the skills, no hierarchy	0.6	0.6	0.6	0.6	0.6	0.6	0.6

Table 24: Probabilities used for the states in which all of the previous levels are also obtained

Case #	Description	p(L1)	p(L2|L1)	p(L3|L2L1)	p(L4|L3L2L1)
1	Equal prob of any ability, standard progression	0.8	0.6	0.4	0.2
2	Mostly high ability, standard progression	0.95	0.85	0.7	0.4
3	Mostly middle ability, standard progression	0.9	0.75	0.45	0.15
4	Mostly low ability, standard progression	0.95	0.65	0.25	0.1
5	Equal prob. of any ability, reversing levels 2 and 3	0.8	0.4	0.6	0.2
6	Mostly middle ability, reversing levels 2 and 3	0.9	0.45	0.75	0.15
7	Equal prob. of having either level 2 or level 3 skills	0.8	0.6	0.6	0.4
8	Equal prob. of having any of the skills	0.6	0.6	0.6	0.6

Table 25: Probability structure for the states when one of the previous levels is not obtained

	P(L2|~L1) P(L3|~L1L2)
	P(L3|~L1~L2)
	P(L4|~L1L2L3)
	P(L4|~L1~L2L3)
	P(L4|~L1L2~L3)
	P(L4|~L1~L2~L3)
		P(L3|L1~L2)	P(L4| L1~L2L3)
			P(L4| L1L2~L3)
				P(L4| L1~L2~L3)

loose hierarchy	0.2	0.2	0.2	0.2
strict hierarchy	0	0	0	0
level 1 loose hierarchy	0.2	=P(L3|L1L2)	=P(L4| L1L2L3)	=P(L4| L1L2L3)
level 1 loose hierarchy, level 4 loosely requires level 2 or level 3	0.2	=P(L3|L1L2)	=P(L4| L1L2L3)	.2

The sample sizes that will be used are 50, 200 and 500. Examining a sample size of 50 might give some insight as to how useful this type of modeling might be for classrooms. After running selected pilot cells it was found that using a sample size of 500 gave comparable results to those with a sample size of 1000 but could be completed in half the time; therefore it was deemed that using 500 would be a large enough sample. An additional sample size of 200 was chosen to use as the medium sample size. Comparing the results from different sample sizes may provide information regarding whether or not a given model is appropriate for the sample size in question or for how large a sample size may need to be to provide adequate results.

For each cell, 100 replications will be used, which is based on the study by Sinharay (2006). (Note that other simulation studies have used fewer replications, as often the time it takes to run these is a factor.) In each replication, data will be simulated according to the model. This data will be used to estimate the parameters for each of the 4 models, and the fit of these models will be compared along with classification rates. For the case where there is one categorical learning progression variable the transition to four latent variables will be fairly straightforward: the student will have the individual latent variables for all levels less than or equal to the student's overall latent variable.

The reverse direction however is not always as straightforward. The problem arises in the case where students have attributes that do not follow the hierarchical model, such as having Level 1 and Level 3 attributes but not Level 2 attributes. In the case where Models 2-4 are used to generate the data there may be students who do not fit nicely into one class in Model 1. There is literature regarding how to handle this situation when the observables do not follow the appropriate pattern (Corcoran, Mosher, & Rogat, 2009). However, in those cases there is still the belief that the students are at one level of the learning progression and there is error in their measurement. In this case there is no level that truly defines the student. Instead the issue may be addressed by determining theoretically what it is we want to say about the student. In the case where the student truly does not have Level 1 attributes, but does have Level 4 attributes, would we really want to say that they are at a higher level of ability or would we want to say they are a complete novice? In the latent class analysis work there have been methods for dealing with intrinsically unscalable subjects. One method is to just put them in their own class (Dayton & Macready, 1980). However, this method would change what it is that class represents. For this study, when this situation arises the student will be placed in the highest level for which they have those attributes and all of the attributes of the lower levels. This follows from the position that if a student is at a given level they should have mastered all of the previous levels.

As discussed in Chapter 2, this study will use MCMC estimation for estimating the probabilities. This study will use the WinBugs program (Spiegelhalter, Thomas, Best & Lunn, 2003) to perform the MCMC estimation. Winbugs was chosen due to its flexibility in the type of models it can support. While this need for flexibility might not be as important as in Study 2, for consistency it was decided that Winbugs would be used for both studies. In Winbugs several chains can be started with different start values. For this study 3 chains will be started (as similar to Levy & Mislevy, 2004 and Almond, Yan, & Hemat, 2008) with starting values at the low end of the distributions, the high end of the starting distributions and the middle range of the distribution.

It is important in MCMC estimation to check for convergence of the chains (Gelman et al., 2004). One statistic that has been used to check the convergence of multiple chains is the Gelman-Rubin statistic (Cowles & Carlin, 1996). This method provides a potential scale reduction factor (PSRF) for each variable in the simulation, as well as a multivariate PSRF (MPSRF) which gives a statistic for all of the variables in the chain (Brooks & Gelman, 1997). A recommendation is that a value of over 1.2 for the MPSRF indicates that the chains have not converged (de la Torre & Douglas, 2004). In this case the simulation will be run for 10000 iterations with a burn-in of 6000. Samples from all cells were checked and it was found that in all cases convergence was obtained using this many iterations.

In addition in MCMC prior distributions are specified for each of the variables. In this study non-informative priors were used because it was desired for the parameter estimates to be minimally influenced by the use of prior information. For the probability of a correct response a beta prior with parameters a=2 and b=2 was used. This implied that there was an equal probability of either a correct or incorrect response but the belief in this prior was not very strong – it is equivalent to the information of two observations, one in each of the two categories. For Model 1 a Dirichlet prior was used with a_i =2 for the probability of being at any of the levels of the learning progression. Similarly a beta prior, again with parameters a=2 and b=2, was used for the latent variable in Models 2-4. This again indicates that the initial belief is that all levels of the learning progression are equally likely but the strength of that belief is not very strong.

Model fit

For this analysis several models will be compared to determine how well each model can be estimated as well as which model may be best suited in different situations. In order to determine if one model outperforms another model there must be a method for comparing these models. Several different methods for examining model fit will be used depending on the question being addressed.

For a simulation study the parameters are known ahead of time and therefore the model can be checked to determine how close the resulting parameters are from the initial parameters. In the Bayesian paradigm, in this case with MCMC estimation, inferences for the parameters are carried out through the posterior distribution. One simple check is to see if the true value of the parameters is within the 95% confidence range of the estimated parameters (Almond, Yan & Hemat, 2008). The number of parameters that are recovered can then be kept track of and averaged across different replications. These averages can be compared across different simulation configurations as well as between different models.

Comparing how well parameters can be recovered does not provide information about which model is the best fit for a particular situation. In order to determine how well the models fit the data, a comparison can be made by examining fit statistics for each model. While methodologies exist for using replicated data based on the posterior predicted model in order to obtain some measure of fit (Gelman, 2003;Levy, Crawford, Fay, & Poole, 2011), these methodologies are still being developed and are not necessarily used as a hypothesis test for overall fit (Levy, Crawford, Fay, & Poole, 2011). In addition this research is concerned with relative fit of the models and therefore leaves absolute model fit to further studies. The issue of absolute fit would be more of concern when comparing across different types of models, as the issue here assumes that BIN have already been picked as the overall model to use and the question is around the issue of which BIN is the best BIN to use. The relative fit of each model will be examined using the likelihood of the data and computing fit statistics.

For each subject in the simulated data set the probability of their given response pattern can be computed. The resulting probability of the entire dataset can also be computed. This results in the likelihood of the data given the parameters of the model. Using this likelihood, different information criteria can be computed. Once these are computed they can be compared across models, with the lower statistic representing the model that is said to fit the best.

One statistic that can be used is Akaike's information criterion (AIC, Burham & Anderson, 2004). When using this with MCMC estimation it can be found by:

$AIC = \overline{D(\xi)} + 2p$ where p is the number of estimated parameters.

In this equation

$D(\xi) = -2\log(p(y \mid \xi) + 2\log(f(y))$

where y is the data, ξ is the posterior mean of the parameters and f(y) is a function of the data alone (and therefore is often not used in the calculation of the statistic, as when comparing two models on the same set of data this part would drop out of the equation) (Spiegelhalter et al., 2002, Li, Cohen, Kim & Cho, 2009).

The AIC has been criticized since it does not take in to account the sample size and so it does not always work as well compared to other fit statistics when the sample size is large (Henson, Reise, & Kim, 2007). It is often compared to the Bayesian inference criterion which is specified as: $BIC = \overline{D(\xi)} + p(\log N)$ where N is the sample size, and p is the number of estimated parameters (Li, Cohen, Kim & Cho, 2009). When the sample size is large AIC tends to indicate better fit (than may be appropriate) for models with more parameters, while the BIC tends to indicate better fit for the models with fewer parameters (Burnham & Anderson, 2004). Therefore, it has been suggested that a good approach is to use both of these statistics in order to determine which model to select (Kuha, 2004).

Another information criterion is the deviance information criterion (DIC). This criterion is built into WinBugs and is designed to be used with MCMC estimation. It uses the definition of the effective number of parameters which is the expected deviance minus the deviance evaluated at the posterior expectations (Spiegelhalter, Best & Carlin, 1998) $p_D = E_{\Theta|y}[D] - D(E_{\Theta|y}[\Theta]) = \overline{D} - D(\overline{\Theta})$ and $DIC = D(\overline{\Theta}) + 2p_D$ (Spiegelhalter, Best & Carlin, 1998).

This study will record the AIC, BIC and DIC for each of the models and then determine which model has the lowest of these values which is an indicator of fit. While it is expected that the generating model should fit the best it will also be of interest to determine if this is indeed the case, and which other models have similar fit.

Classification Accuracy

In the case of learning progressions what may be of most interest is the resulting classification of students into levels of the learning progressions. While the actual structural parameter values may not be of great importance, misclassification of students could lead to them being placed at a higher level than they are in which case they may struggle to learn the material or at a lower level which would then lead them to repeat information they may already know.

In the study of classification two data sets were generated from the same set of starting parameters. They distinguish classification accuracy from a run which classification is carried out with structural parameters estimated from the same data set, and the other in which the structural parameters were estimated from one data set and used to carry out classification on a new set of students from the same population. One data set was used to run the model and determine the resulting parameters. These parameters were then used to classify the students in this data set. Additionally, the second data set was used to check how well the resulting model was able to classify students. This second data set was used to determine if there is a drop off in the rate of classification when a different sample of students is used.

In order to determine the classification accuracy, the percent of students correctly classified was recorded. In addition the adjusted Rand index (Steinley, 2004) was used as an indicator of how well the classifications from each model match the original classifications.

In the case where the generating model was Models 2-4 but the model being estimated is Model 1 there may be students who do not fit into one of the levels of the LP in Model 1. For example, if a student has the attributes of Level 2 and Level 4 but not Level 1 or Level 3 then there is no corresponding level of the learning progression that captures that behavior. In this case the student will be labeled as misclassified regardless of which level of the learning progression they are assigned, because there is no class that truly represents their ability structure.

Each of the three presented methods (parameter recovery, relative model fit, and classification accuracy) answers a slightly different question and when combined should provide support for the benefit or drawback of using each of the models in the given situation. Therefore it was deemed that these methods would be appropriate for this study. These model fit indices will be calculated for each replication, and then compared across cells to determine how well each model performed for the different possible parameter distributions.

Results

In general the parameters were able to be recovered on a fairly consistent basis for all models (see Table 26), with all cells recovering (on average across all repetitions of the cell) at least 90% of the parameters correctly. For each cell in the model parameter recovery was determined by how well the model that was used to generate the data was able to recover the parameters (and did not consider how well a model that was not used to generate the parameters was able to recover parameters).. The percent of parameters recovered (using the 95% central interval described above) for the probabilities associated with the latent class was within one standard deviation of each other for Models 1-3, but Model 4 was about 2 standard deviations below that of the other models.

Examining the individual variable probabilities recovered showed that Models 2-4 were all within one standard deviation from each other, but more than 5 standard deviations below Model 1. One possible explanation for this is that the probability of students who have the lower level attributes tended to be high so there was not as many examples of students who did not have these attributes getting the answers correct when they did not have the appropriate attribute which could cause the estimation of those parameters to be incorrect.

Table 26: Parameter Recovery information, averaged across all cells for each Model

	% LP probs recovered across all cells				% Obs probs recovered across all cells			
Model	min	max	average	std. dev.	min	max	average	std. dev.
1	0.94	0.97	0.954	0.01	0.951	0.967	0.958	0.004
2	0.933	0.98	0.955	0.013	0.913	0.958	0.934	0.015
3	0.939	0.983	0.959	0.011	0.911	0.964	0.933	0.015
4	0.908	0.993	0.956	0.013	0.913	0.961	0.935	0.014

In terms of model fit, Models 1 and 3 seemed to be the best fitting models, regardless of which model was used to generate the data. In general the DIC favored Model 1 particularly as the sample size increased, which may be expected as the DIC penalizes more complex models (Wheeler, Hickson & Waller, 2010). The one exception to that was the case where Model 4 was the generating model and in this case the DIC picked Model 3 as the model with the best fit instead.

When Model 1 was the generating model and the sample size was low then the BIC picked Model 3 as the best fit and the AIC picked Model 3. When the sample size was high then while the BIC still picked Model 3, the AIC was mixed between Model 1 and Model 3 (see

Table 27). In particular it seemed to favor Model 3 when there was an unequal distribution of students along ability range and the probability of a correct response stayed the same as the ability level increased over the requirements for the item. It could be that since Model 1 is a special case of Model 3 and since the probability of a correct response is the same if a student has the required attribute regardless of if a student has higher level attributes that Model 3 is able to capture this structure as well or better than Model 1. For the small sample size, the DIC indicated Model 1 when there was an equal probability of students in each class and Model 3 otherwise, but as sample size increased it indicated that Model 1 was the best fitting model.

Table 27: Proportion of replications in which each model was picked as the best fit for data generated by Model 1

SS	LP	OV	% of times chosen by AIC				% of times chosen by BIC				% of times chosen by DIC			
			M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	1	1	0	0	**1**	0	0	0	**1**	0	**0.63**	0	0.37	0
50	1	2	0	0	**1**	0	0	0	**1**	0	**0.54**	0	0.46	0
50	2	1	0.01	0	**0.99**	0	0	0	**1**	0	0.35	0	**0.65**	0
50	2	2	0	0	**1**	0	0	0	**1**	0	0.25	0	**0.75**	0
50	3	1	0	0	**1**	0	0	0	**1**	0	0.32	0	**0.68**	0
50	3	2	0	0	**1**	0	0	0	**1**	0	0.33	0	**0.67**	0
50	4	1	0	0	**1**	0	0	0.05	**0.95**	0	0.15	0	**0.85**	0
50	4	2	0	0	**1**	0	0	0.01	**0.99**	0	0.2	0	**0.8**	0
200	1	1	**0.68**	0	0.32	0	0.15	0	**0.85**	0	**1**	0	0	0
200	1	2	0.31	0	**0.69**	0	0.05	0	**0.95**	0	**0.99**	0	0.01	0
200	2	1	0.16	0	**0.84**	0	0.1	0	**0.9**	0	**0.97**	0	0.03	0
200	2	2	0.04	0	**0.96**	0	0	0	**1**	0	0.38	0	**0.62**	0
200	3	1	0.19	0	**0.81**	0	0.17	0	**0.83**	0	**1**	0	0	0
200	3	2	0.09	0	**0.91**	0	0.02	0	**0.98**	0	**0.79**	0	0.21	0
200	4	1	0.03	0	**0.97**	0	0.02	0	**0.98**	0	**0.92**	0	0.08	0
200	4	2	0.04	0	**0.96**	0	0	0	**1**	0	0.42	0	**0.58**	0
500	1	1	**1**	0	0	0	0.3	0	**0.7**	0	**1**	0	0	0
500	1	2	**0.99**	0	0.01	0	0.21	0	**0.79**	0	**1**	0	0	0
500	2	1	**0.9**	0	0.1	0	0.04	0	**0.96**	0	**1**	0	0	0
500	2	2	0.07	0	**0.93**	0	0.06	0	**0.94**	0	**0.87**	0	0.13	0
500	3	1	**0.99**	0	0.01	0	0.1	0	**0.9**	0	**1**	0	0	0
500	3	2	0.25	0	**0.75**	0	0.07	0	**0.93**	0	**1**	0	0	0
500	4	1	**0.57**	0	0.43	0	0.07	0	**0.93**	0	**1**	0	0	0
500	4	2	0.04	0	**0.96**	0	0.04	0	**0.96**	0	**0.84**	0	0.16	0

When Model 2 was the generating model the DIC picked Model 1 as the best fitting model in all cases. The AIC and the BIC statistics picked Model 2 as the best fitting model most of the time when the sample was small (see Table 28). As the sample size increased, the AIC started picking Model 1 as the best fitting model for all cases, while BIC indicated Model 1 only in some cases, particularly those cases where a standard progression of attributes was not necessarily followed.

This result seems surprising as one of the theorized reasons for choosing Model 2 would be to allow for the attributes of the LP to not follow a standard progression. However, this indicates that Model 1 would be the best fitting model in these situations, which provides justification for the use of Model 1 even in situations for which the relationship between the attributes differ than that specified in Model 1. An examination of select cells showed that the difference between the fit values was more than 10, which is one rule of thumb when selecting models (Burnham & Anderson, 2004).

AIC tended to indicate Model 3 was the best fitting model when the data was generated using Model 3 and the sample sizes were small (see Table 29). As the sample sizes increased the AIC picked Model 1 as the best fitting model, except in the cases where the data should have followed a strict hierarchy and the students were not equally distributed in ability levels, in which case Model 3 was picked. For small sample sizes the BIC indicated that either Model 2 or Model 3 was the best fitting model while as sample sizes increased either Model 1 or Model 3 was chosen. Interesting was that in the case where a strict hierarchy was followed Model 3 tended to be chosen over Model 1. The DIC indicated Model 1for large sample sizes and either Model 3 or Model 1 for small sample sizes.

Table 28: Proportion of replications in which each model was picked as the best fit for data generated using Model 2

SS	LP	OV	% of times chosen by AIC				% of times chosen by BIC				% of times chosen by DIC			
			M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	1	1	0.02	**0.87**	0.04	0.07	0	**1**	0	0	**0.81**	0.15	0.03	0.01
50	2	1	0	**0.91**	0.03	0.06	0	**1**	0	0	0.31	**0.53**	0.05	0.11
50	3	1	0.02	**0.96**	0.01	0.01	0	**1**	0	0	**0.58**	0.33	0.08	0.01
50	4	1	0	**0.98**	0.01	0.01	0	**1**	0	0	**0.59**	0.33	0.08	0
50	5	1	0.01	**0.91**	0.05	0.03	0	**1**	0	0	**0.83**	0.14	0.02	0.01
50	6	1	0.01	**0.93**	0.02	0.04	0	**1**	0	0	**0.72**	0.21	0.06	0.01
50	7	1	0.02	**0.71**	0	0.27	0	**0.96**	0	0.04	**0.76**	0.13	0.02	0.09
50	8	1	0.12	**0.6**	0.04	0.24	0	**0.96**	0	0.04	**0.86**	0.08	0.04	0.02
200	1	1	**0.86**	0.13	0	0.01	0.09	**0.9**	0	0.01	**1**	0	0	0
200	2	1	0.16	**0.77**	0.01	0.06	0.01	**0.99**	0	0	**0.72**	0.22	0.03	0.03
200	3	1	0.45	**0.53**	0	0.02	0.05	**0.95**	0	0	**0.98**	0.02	0	0
200	4	1	0.38	**0.58**	0.04	0	0.05	**0.95**	0	0	**0.97**	0.02	0.01	0
200	5	1	**0.96**	0.04	0	0	0.25	**0.75**	0	0	**1**	0	0	0
200	6	1	**0.73**	0.27	0	0	0.03	**0.97**	0	0	**1**	0	0	0
200	7	1	**0.87**	0.13	0	0	0.15	**0.84**	0	0.01	**0.99**	0	0	0.01
200	8	1	**0.99**	0.01	0	0	0.39	**0.6**	0	0.01	**1**	0	0	0
500	1	1	**1**	0	0	0	**0.9**	0.1	0	0	**1**	0	0	0
500	2	1	**0.8**	0.17	0	0.03	0.03	**0.97**	0	0	**1**	0	0	0
500	3	1	**0.99**	0.01	0	0	0.18	**0.82**	0	0	**1**	0	0	0
500	4	1	**0.99**	0	0.01	0	0.06	**0.94**	0	0	**1**	0	0	0
500	5	1	**1**	0	0	0	**0.95**	0.05	0	0	**1**	0	0	0
500	6	1	**1**	0	0	0	0.47	**0.53**	0	0	**1**	0	0	0
500	7	1	**1**	0	0	0	**0.8**	0.2	0	0	**1**	0	0	0
500	8	1	**1**	0	0	0	**1**	0	0	0	**1**	0	0	0

Table 29: Proportion of replications in which each model was picked as the best fit for data generated using Model 3

SS	LP	OV	% of times chosen by AIC				% of times chosen by BIC				% of times chosen by DIC			
			M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	1	1	0	0	**1**	0	0	0.02	**0.98**	0	**0.56**	0	0.44	0
50	2	1	0.03	0.34	**0.4**	0.23	0	**0.7**	0.15	0.15	**0.75**	0.05	0.18	0.02
50	3	1	0	0	**1**	0	0	0	**1**	0	0.16	0	**0.84**	0
50	4	1	0.01	0.06	**0.55**	0.38	0	0.16	**0.52**	0.32	0.34	0.03	**0.48**	0.15
50	5	1	0	0.03	**0.97**	0	0	0.09	**0.91**	0	0.43	0	**0.57**	0
50	6	1	0.01	0.22	**0.53**	0.24	0	**0.63**	0.28	0.09	**0.65**	0.01	0.25	0.09
50	7	1	0.01	0.36	**0.63**	0	0	**0.77**	0.23	0	0.27	0	**0.73**	0
50	8	1	0.01	**0.83**	0.08	0.08	0	**0.96**	0.02	0.02	**0.71**	0.18	0.09	0.02
50	9	1	0.05	0.23	**0.45**	0.27	0	**0.54**	0.26	0.2	**0.8**	0	0.19	0.01
50	10	1	0.04	**0.75**	0.09	0.12	0	**0.96**	0.01	0.03	**0.7**	0.13	0.17	0
50	11	1	0	0.1	**0.58**	0.32	0	0.29	**0.46**	0.25	**0.68**	0	0.28	0.04
50	12	1	0.04	0.36	0.11	**0.49**	0	**0.82**	0.03	0.15	**0.73**	0.04	0.16	0.07
50	13	1	0.04	0.02	**0.49**	0.45	0	0.3	**0.38**	0.32	**0.78**	0	0.18	0.04
50	14	1	0.03	0.4	0.1	**0.47**	0	**0.8**	0.02	0.18	**0.71**	0.07	0.13	0.09
50	15	1	0.08	0	0.43	**0.49**	0	0.02	**0.46**	0.52	**0.88**	0	0.09	0.03
50	16	1	0.11	**0.68**	0.01	0.2	0	**0.92**	0	0.08	**0.82**	0.12	0.02	0.04
200	1	1	0.27	0	**0.73**	0	0.02	0	**0.98**	0	**0.98**	0	0.02	0
200	2	1	**0.95**	0	0.05	0	0.25	0.05	**0.47**	0.23	**0.99**	0	0.01	0
200	3	1	0.05	0	**0.95**	0	0.01	0	**0.99**	0	**0.5**	0	**0.5**	0
200	4	1	0.09	0	**0.49**	0.42	0.01	0	**0.5**	0.49	**0.73**	0	0.24	0.03
200	5	1	0.12	0	**0.88**	0	0.01	0	**0.99**	0	**0.78**	0	0.22	0
200	6	1	**0.63**	0	0.28	0.09	0.07	0.03	**0.63**	0.27	**0.98**	0	0.02	0
200	7	1	0.03	0	**0.97**	0	0.01	0	**0.99**	0	0.47	0	**0.53**	0
200	8	1	**0.57**	0.07	0.25	0.11	0.02	**0.76**	0.19	0.03	**0.97**	0	0.03	0
200	9	1	**0.94**	0	0.05	0.01	0.27	0.01	**0.5**	0.22	**1**	0	0	0
200	10	1	**0.74**	0.05	0.12	0.09	0.1	**0.72**	0.06	0.12	**0.99**	0	0.01	0
200	11	1	**0.74**	0	0.19	0.07	0.09	0	**0.67**	0.24	**0.99**	0	0.01	0
200	12	1	**0.91**	0	0.07	0.02	0.21	0.21	0.22	**0.36**	**0.99**	0	0.01	0
200	13	1	**0.89**	0	0.03	0.08	0.21	0	**0.38**	0.41	**1**	0	0	0
200	14	1	**0.83**	0	0.08	0.09	0.17	0.16	0.22	**0.45**	**1**	0	0	0
200	15	1	**0.98**	0	0	0.02	0.33	0	**0.37**	0.3	**1**	0	0	0
200	16	1	**1**	0	0	0	0.43	**0.56**	0	0.01	**1**	0	0	0

SS	LP	OV	% of times chosen by AIC				% of times chosen by BIC				% of times chosen by DIC			
			M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
500	1	1	**0.99**	0	0.01	0	0.19	0	**0.81**	0	**1**	0	0	0
500	2	1	**1**	0	0	0	**0.96**	0	0.02	0.02	**1**	0	0	0
500	3	1	0.06	0	**0.94**	0	0.06	0	**0.94**	0	**0.87**	0	0.13	0
500	4	1	**0.55**	0	0.22	0.23	0.03	0	0.46	**0.51**	**0.98**	0	0.02	0
500	5	1	0.26	0	**0.74**	0	0.08	0	**0.92**	0	**1**	0	0	0
500	6	1	**1**	0	0	0	0.35	0	**0.38**	0.27	**1**	0	0	0
500	7	1	0.02	0	**0.98**	0	0.02	0	**0.98**	0	**0.86**	0	0.14	0
500	8	1	**1**	0	0	0	0.22	0.11	**0.53**	0.14	**1**	0	0	0
500	9	1	**1**	0	0	0	**0.88**	0	0.07	0.05	**1**	0	0	0
500	10	1	**1**	0	0	0	**0.47**	0.1	0.34	0.09	**1**	0	0	0
500	11	1	**1**	0	0	0	**0.47**	0	0.4	0.13	**1**	0	0	0
500	12	1	**1**	0	0	0	**0.89**	0	0.06	0.05	**1**	0	0	0
500	13	1	**1**	0	0	0	**0.83**	0	0.11	0.06	**1**	0	0	0
500	14	1	**1**	0	0	0	**0.93**	0	0.04	0.03	**1**	0	0	0
500	15	1	**1**	0	0	0	**1**	0	0	0	**1**	0	0	0
500	16	1	**1**	0	0	0	**1**	0	0	0	**1**	0	0	0

In general, Model 4 was not picked to be the best fitting model even when it was the generating model (see Table 30). For small sample sizes AIC and BIC tended to indicate Model 2 when there were mostly middle or low ability students, Model 4 when the probability was equal across all of the attributes, and Model 3 otherwise. As the sample size increased the AIC indicated Model 1 was a better fit except in the cases where there was a strict hierarchy and the distribution of students was skewed. The BIC indicated Model 1 was the best fitting model when the ability levels of the students were equally distributed and was split between the rest of the models otherwise, although Model 2 seemed to be picked more often when the ability level of the students was low and Model 4 was picked when there were mostly high ability students. The DIC tended to pick Model 1 across sample size.

Table 30: Proportion of replications in which each model was picked as the best fit for data generated using Model 4

SS	LP	OV	% of times chosen by AIC				% of times chosen by BIC				% of times chosen by DIC			
			M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	1	1	0.01	0	**0.99**	0	0	0.01	**0.99**	0	**0.58**	0	0.42	0
50	2	1	0.03	**0.43**	0.34	0.2	0	**0.76**	0.15	0.09	**0.81**	0.04	0.12	0.03
50	3	1	0.04	**0.44**	0.29	0.23	0	**0.74**	0.13	0.13	**0.9**	0.03	0.05	0.02
50	4	1	0.03	**0.47**	0.32	0.18	0	**0.81**	0.1	0.09	**0.75**	0.06	0.18	0.01
50	5	1	0	0	**1**	0	0	0	**1**	0	0.16	0	**0.84**	0
50	6	1	0	0.11	**0.53**	0.36	0	0.36	**0.42**	0.22	0.39	0.01	**0.42**	0.18
50	7	1	0	0.46	0.05	**0.49**	0	**0.75**	0.03	0.22	**0.45**	0.13	0.16	0.26
50	8	1	0	0.12	**0.55**	0.33	0	0.35	**0.45**	0.2	0.34	0	**0.47**	0.19
50	9	1	0	0	**1**	0	0	0.06	**0.94**	0	0.37	0	**0.63**	0
50	10	1	0	0.43	**0.47**	0.1	0	**0.81**	0.16	0.03	**0.66**	0.05	0.26	0.03
50	11	1	0.01	**0.5**	0.29	0.2	0	**0.76**	0.14	0.1	**0.62**	0.11	0.23	0.04
50	12	1	0.01	0.36	**0.42**	0.21	0	**0.77**	0.13	0.1	**0.65**	0.06	0.24	0.05
50	13	1	0	0.36	**0.64**	0	0	**0.79**	0.21	0	0.26	0.05	**0.69**	0
50	14	1	0	**0.91**	0.04	0.05	0	**1**	0	0	**0.64**	0.25	0.07	0.04
50	15	1	0.01	**0.93**	0.04	0.02	0	**1**	0	0	**0.68**	0.2	0.1	0.02
50	16	1	0	**0.9**	0.03	0.07	0	**1**	0	0	**0.67**	0.22	0.08	0.03
50	17	1	0.06	**0.59**	0.02	0.33	0	**0.88**	0	0.12	**0.87**	0.08	0.02	0.03
50	18	1	0.03	0.18	**0.5**	0.29	0	**0.54**	0.28	0.18	**0.83**	0.01	0.14	0.02
50	19	1	0.01	**0.66**	0.04	0.29	0	**0.93**	0	0.07	**0.79**	0.07	0.02	0.12
50	20	1	0	0.09	**0.72**	0.19	0	0.29	**0.55**	0.16	**0.66**	0	0.3	0.04
50	21	1	0.05	0.14	0.07	**0.74**	0	**0.53**	0.03	0.44	**0.85**	0.03	0.04	0.08
50	22	1	0.02	0.06	**0.5**	0.42	0	0.26	**0.39**	0.35	**0.77**	0	0.19	0.04
50	23	1	0.19	0	0	**0.81**	0	0.03	0	**0.97**	**0.93**	0	0	0.07
50	24	1	0.03	0.01	0.38	**0.58**	0	0.07	0.34	**0.59**	**0.77**	0	0.16	0.07

SS	LP	OV	% of times chosen by AIC				% of times chosen by BIC				% of times chosen by DIC			
			M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
200	1	1	0.42	0	**0.58**	0	0.05	0	**0.95**	0	**1**	0	0	0
200	2	1	**0.97**	0	0.03	0	0.14	0.06	0.34	**0.46**	**1**	0	0	0
200	3	1	**0.91**	0	0.02	0.07	0.19	0.12	0.09	**0.6**	**0.99**	0	0.01	0
200	4	1	**0.92**	0	0.04	0.04	0.25	0.01	**0.4**	0.34	**1**	0	0	0
200	5	1	0.01	0	**0.99**	0	0	0	**1**	0	0.44	0	**0.56**	0
200	6	1	0.15	0	0.26	**0.59**	0.01	0	0.29	**0.7**	**0.93**	0	0.03	0.04
200	7	1	0.26	0.01	0	**0.73**	0.07	0.1	0	**0.83**	**0.82**	0	0.04	0.14
200	8	1	0.22	0	**0.42**	0.36	0.02	0	**0.48**	**0.5**	**0.75**	0	0.22	0.03
200	9	1	0.11	0	**0.89**	0	0.02	0	**0.98**	0	**0.83**	0	0.17	0
200	10	1	**0.61**	0	0.25	0.14	0.08	0.03	**0.61**	0.28	**0.97**	0	0.03	0
200	11	1	**0.57**	0	0.14	0.29	0.06	0.13	0.21	**0.6**	**0.99**	0	0	0.01
200	12	1	**0.59**	0	0.31	0.1	0.04	0.02	**0.76**	0.18	**0.99**	0	0.01	0
200	13	1	0.08	0	**0.92**	0	0	0	**1**	0	0.45	0	**0.55**	0
200	14	1	**0.51**	0.21	0.25	0.03	0.05	**0.87**	0.08	0	**0.99**	0	0.01	0
200	15	1	**0.42**	0.3	0.22	0.06	0.03	**0.85**	0.07	0.05	**0.98**	0	0.02	0
200	16	1	**0.37**	0.25	0.32	0.06	0.02	**0.89**	0.09	0	**0.97**	0.02	0.01	0
200	17	1	**0.92**	0	0	0.08	0.18	0.08	0	**0.74**	**1**	0	0	0
200	18	1	**0.94**	0	0.04	0.02	0.22	0	**0.43**	0.35	**1**	0	0	0
200	19	1	**0.75**	0.02	0	0.23	0.16	0.2	0	**0.64**	**0.99**	0	0	0.01
200	20	1	**0.62**	0	0.29	0.09	0.11	0	**0.66**	0.23	**1**	0	0	0
200	21	1	**0.94**	0	0	0.06	0.18	0	0	**0.82**	**1**	0	0	0
200	22	1	**0.87**	0	0.08	0.05	0.15	0	0.35	**0.5**	**1**	0	0	0
200	23	1	**0.99**	0	0	0.01	**0.57**	0	0	0.43	**1**	0	0	0
200	24	1	**0.99**	0	0.01	0	0.36	0	0.16	**0.48**	**1**	0	0	0

SS	LP	OV	% of times chosen by AIC				% of times chosen by BIC				% of times chosen by DIC			
			M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
500	1	1	**0.99**	0	0.01	0	0.21	0	**0.79**	0	**1**	0	0	0
500	2	1	**1**	0	0	0	**0.94**	0	0.02	0.04	**1**	0	0	0
500	3	1	**1**	0	0	0	**0.95**	0	0	0.05	**1**	0	0	0
500	4	1	**1**	0	0	0	**0.91**	0	0.08	0.01	**1**	0	0	0
500	5	1	0.02	0	**0.98**	0	0.02	0	**0.98**	0	**0.85**	0	0.15	0
500	6	1	**0.91**	0	0	0.09	0.07	0	0.03	**0.9**	**1**	0	0	0
500	7	1	**0.87**	0	0	0.13	0.02	0	0	**0.98**	**0.99**	0	0	0.01
500	8	1	**0.81**	0	0.12	0.07	0.08	0	**0.48**	0.44	**1**	0	0	0
500	9	1	0.22	0	**0.78**	0	0.06	0	**0.94**	0	**1**	0	0	0
500	10	1	**1**	0	0	0	0.33	0	0.28	**0.39**	**1**	0	0	0
500	11	1	**1**	0	0	0	0.29	0	0.07	**0.64**	**1**	0	0	0
500	12	1	**1**	0	0	0	0.29	0	**0.54**	0.17	**1**	0	0	0
500	13	1	0.05	0	**0.95**	0	0.05	0	**0.95**	0	**0.9**	0	0.1	0
500	14	1	**1**	0	0	0	0.21	**0.42**	0.33	0.04	**1**	0	0	0
500	15	1	**1**	0	0	0	0.13	**0.43**	0.34	0.1	**1**	0	0	0
500	16	1	**1**	0	0	0	0.12	0.41	**0.37**	0.1	**1**	0	0	0
500	17	1	**1**	0	0	0	**0.96**	0	0	0.04	**1**	0	0	0
500	18	1	**1**	0	0	0	**0.94**	0	0.03	0.03	**1**	0	0	0
500	19	1	**1**	0	0	0	**0.56**	0	0	0.44	**1**	0	0	0
500	20	1	**0.99**	0	0.01	0	**0.43**	0	**0.44**	0.13	**1**	0	0	0
500	21	1	**1**	0	0	0	**0.87**	0	0	0.13	**1**	0	0	0
500	22	1	**1**	0	0	0	**0.81**	0	0.06	0.13	**1**	0	0	0
500	23	1	**1**	0	0	0	**1**	0	0	0	**1**	0	0	0
500	24	1	**1**	0	0	0	**1**	0	0	0	**1**	0	0	0

Overall, it seems that Model 1 tends to be the best fitting model when the sample size is large, Model 2 might be as good or better when the students are mostly low ability students, and Model 3 might be a good model for the rest of the cases. While in some cases Model 4 may be a good model if the ability levels of the students are high, it does not generally seem to be the best fitting model.

In terms of classification, when the same data that was used for generating the parameters of the model was used to classify the students then in all cases Model 1 had a higher classification rate (see Table 31) although no model performed poorly. When a separate sample was used, overall classification dropped slightly (as would be expected due to differences in the sample that generated the parameters and the sample that was used to test the model).

Table 31: Average percent of students classified correctly across all cells for the data that was used to estimate the model

Generated by	Classified by M1		Classified by M2		Classified by M3		Classified by M4	
	Ave.	St. Dev.	Ave.	St. Dev.	Ave.	St. Dev	Ave.	St. Dev.
Model 1	84.40%	2.40%	72.30%	3.50%	81.10%	1.50%	80.60%	1.50%
Model 2	80.10%	3.40%	68.10%	4.60%	68.30%	4.50%	68.50%	4.40%
Model 3	79.10%	1.90%	66.80%	4.70%	69.00%	6.20%	69.10%	6.10%
Model 4	79.10%	1.60%	68.40%	4.50%	70.50%	5.50%	70.90%	5.10%

When Model 1 was the generating model and the same data was used to generate the parameters as to classify the students then Model 1 had a higher classification rate and the adjusted Rand index was largest for Model 1 (see Table 33).

The Rand index is a number between 0 and 1 where numbers closest to 1 indicate a higher correspondence. The answer space is first partitioned into different pieces such that no piece overlaps and the combination of pieces cover the entire answer space (i.e. for a categorical answer space this could be that each partition is one category). A matrix is set up such where the rows and the columns are the partition and the cells are the total number of cases that occur in each partition. For example, in Table 32, the cell t_{12} would be the number of times a student who was actually at level 1 was classified to be at level 2.

Table 32: Adjusted Rand partition table

group	q_1	q_2	…	q_C	Total
p_1	t_{11}	t_{12}	…	t_{1C}	t_{1+}
p_2	t_{21}	t_{22}	…	t_{2C}	t_{2+}
…	…	…	..	…	…
p_R	t_{R1}	t_{R2}	…	t_{RC}	t_{R+}
Total	t_{+1}	t_{+2}	…	t_{+C}	N

The Rand index then generates numbers that represent the different types of pairs as follows:

$$a = \frac{\sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^{2} - N}{2}$$

$$b = \frac{\sum_{r=1}^{R} t_{r+}^{2} - \sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^{2}}{2}$$

$$c = \frac{\sum_{c=1}^{C} t_{+c}^{2} - \sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^{2}}{2}$$

$$d = \frac{\sum_{r=11}^{R}\sum_{c=1}^{C} t_{rc}^{2} + N^{2} - \sum_{r=1}^{R} t_{r+}^{2} - \sum_{c=1}^{C} t_{+c}^{2}}{2}$$

The adjusted Rand index is then found by using:

$$\frac{\left(\left(\frac{N!}{(N-2)!2!}\right)(a+d) - [(a+b)(a+c)+(c+d)(b+d)]\right)}{\left(\left(\frac{N!}{(N-2)!2!}\right)^{2} - [(a+b)(a+c)+(c+d)(b+d)]\right)}$$

For this study, the resulting level of the student was compared to the generated (known true) level of the student. For Models 2-4 students could have 16 possible outcomes for their level (they had two possibilities for each of the four level variables), so a number from 1-16 was assigned based on their starting level and this was compared to their estimated level. The level of the learning progression for Model 1 could also be converted to one of these numbers by assuming that they have all of the attribute variables for the level they were assigned and all of the previous levels.

Table 33: Classification information when Model 1 generated the data same data was used for classification

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
50	1	1	**87.98%**	70.90%	80.80%	79.74%	**0.726**	0.501	0.606	0.599
50	1	2	**87.60%**	67.78%	79.80%	78.72%	**0.715**	0.472	0.587	0.581
50	2	1	**89.10%**	77.28%	84.30%	83.50%	**0.729**	0.607	0.666	0.667
50	2	2	**87.88%**	74.34%	82.48%	81.52%	**0.713**	0.562	0.629	0.63
50	3	1	**86.86%**	75.48%	81.56%	80.58%	**0.688**	0.555	0.602	0.599
50	3	2	**86.04%**	72.62%	80.18%	78.82%	**0.664**	0.513	0.568	0.559
50	4	1	**87.38%**	77.66%	83.30%	82.20%	**0.693**	0.569	0.621	0.616
50	4	2	**86.54%**	74.62%	81.76%	80.52%	**0.676**	0.545	0.605	0.596
200	1	1	**83.52%**	69.11%	80.02%	79.74%	**0.643**	0.475	0.589	0.588
200	1	2	**82.29%**	65.13%	78.79%	78.46%	**0.616**	0.43	0.562	0.56
200	2	1	**85.32%**	75.81%	83.07%	82.91%	**0.674**	0.593	0.649	0.65
200	2	2	**83.34%**	71.73%	81.73%	81.45%	**0.635**	0.532	0.618	0.617
200	3	1	**82.52%**	72.50%	79.83%	79.67%	**0.6**	0.506	0.564	0.564
200	3	2	**82.21%**	70.12%	79.91%	79.63%	**0.595**	0.489	0.563	0.563
200	4	1	**84.15%**	75.45%	82.13%	81.92%	**0.638**	0.547	0.605	0.604
200	4	2	**82.54%**	73.32%	80.83%	80.35%	**0.606**	0.522	0.581	0.576
500	1	1	**82.71%**	68.86%	79.54%	79.33%	**0.625**	0.47	0.578	0.577
500	1	2	**81.88%**	65.01%	79.34%	79.11%	**0.607**	0.429	0.567	0.566
500	2	1	**83.53%**	74.93%	82.50%	82.44%	**0.642**	0.575	0.632	0.632
500	2	2	**83.23%**	72.07%	82.54%	82.40%	**0.637**	0.542	0.631	0.631
500	3	1	**81.71%**	71.89%	80.19%	79.99%	**0.586**	0.498	0.566	0.565
500	3	2	**80.92%**	69.07%	79.61%	79.26%	**0.573**	0.472	0.554	0.553
500	4	1	**83.14%**	74.85%	81.92%	81.76%	**0.618**	0.535	0.598	0.597
500	4	2	**82.14%**	73.67%	81.24%	81.05%	**0.601**	0.527	0.588	0.588

When Model 1 was the generating model and a separate data set was used to test classification, then for small sample sizes Model 3 had the highest classification rates. With a sample size of 200, Model 1 had the highest classification rates when there was an even distribution in the spread of the student’s ability while Model 3 had a higher classification rate for all other cases. With the large sample size, Model 1 had the highest classification rates except when the students were mostly high ability students, in which case Model 3 had the highest classification rate (see Table 34). For the most part, the adjusted Rand index followed this same pattern. A few exceptions occurred when the sample size was 200, in which case the adjusted Rand index indicated that Model 4 was slightly more consistent with the original classifications.

Table 34: Classification information when Model 1 generated the data and a separate data set was used for classification

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
50	1	1	74.50%	67.60%	**77.60%**	76.30%	0.495	0.46	**0.551**	0.543
50	1	2	72.40%	63.60%	**75.20%**	73.50%	0.461	0.416	**0.514**	0.502
50	2	1	76.80%	73.40%	**80.30%**	78.80%	0.558	0.557	**0.616**	0.608
50	2	2	74.00%	70.10%	**79.20%**	77.80%	0.507	0.509	**0.591**	0.59
50	3	1	72.90%	70.50%	**76.60%**	74.70%	0.449	0.476	**0.523**	0.513
50	3	2	71.80%	67.70%	**75.70%**	74.20%	0.433	0.451	**0.511**	0.505
50	4	1	74.00%	72.60%	**78.30%**	77.10%	0.479	0.501	**0.546**	0.538
50	4	2	73.80%	70.70%	**77.30%**	76.10%	0.477	0.482	**0.535**	0.528
200	1	1	**80.40%**	67.80%	79.10%	78.90%	**0.585**	0.457	0.576	0.575
200	1	2	**78.80%**	64.20%	77.50%	77.30%	**0.553**	0.415	0.539	0.539
200	2	1	81.00%	74.70%	**82.20%**	82.00%	0.61	0.579	0.634	**0.635**
200	2	2	80.40%	71.00%	**81.30%**	80.90%	0.599	0.525	**0.615**	0.614
200	3	1	79.10%	71.50%	**79.30%**	79.10%	0.542	0.495	0.557	**0.558**
200	3	2	77.90%	67.90%	**78.00%**	77.90%	0.522	0.454	0.53	**0.531**
200	4	1	80.30%	73.80%	**80.80%**	80.50%	0.57	0.518	**0.582**	0.579
200	4	2	79.00%	72.00%	**79.70%**	79.40%	0.549	0.501	**0.563**	0.561
500	1	1	**81.60%**	67.90%	79.10%	78.80%	**0.604**	0.457	0.569	0.567
500	1	2	**80.70%**	63.90%	78.40%	78.10%	**0.585**	0.414	0.551	0.55
500	2	1	82.10%	74.50%	**82.40%**	82.30%	0.621	0.569	0.626	**0.627**
500	2	2	81.80%	71.90%	**81.90%**	81.80%	0.618	0.535	**0.623**	0.622
500	3	1	**80.30%**	71.20%	79.80%	79.60%	**0.562**	0.486	0.559	0.558
500	3	2	**79.10%**	68.00%	78.70%	78.40%	**0.541**	0.459	0.54	0.539
500	4	1	**81.70%**	74.20%	81.40%	81.20%	**0.597**	0.527	0.59	0.589
500	4	2	**80.40%**	72.90%	**80.40%**	80.20%	**0.574**	0.517	**0.574**	0.573

When the data was generated by Model 2 and the same data set was used for classification then Model 1 always had the highest classification rate (see Table 35). In addition, the adjusted Rand index was highest for Model 1.

Table 35: Classification information when Model 2 generated the data and the same data was used for classification

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
50	1	1	**83.20%**	66.80%	67.30%	67.80%	**0.613**	0.454	0.462	0.468
50	2	1	**83.10%**	75.30%	75.20%	75.30%	**0.614**	0.544	0.547	0.546
50	3	1	**83.10%**	72.50%	72.50%	72.80%	**0.605**	0.503	0.505	0.51
50	4	1	**86.30%**	75.90%	76.40%	75.90%	**0.654**	0.544	0.555	0.554
50	5	1	**85.50%**	68.30%	68.90%	68.90%	**0.671**	0.483	0.491	0.495
50	6	1	**85.90%**	71.30%	71.70%	72.00%	**0.665**	0.498	0.503	0.51
50	7	1	**82.20%**	65.00%	65.60%	66.80%	**0.603**	0.426	0.433	0.447
50	8	1	**83.80%**	60.80%	61.90%	63.30%	**0.642**	0.374	0.384	0.406
200	1	1	**77.70%**	65.50%	65.60%	65.80%	**0.517**	0.439	0.44	0.442
200	2	1	**77.80%**	71.80%	72.00%	72.20%	**0.516**	0.499	0.502	0.503
200	3	1	**78.10%**	69.90%	70.00%	70.30%	**0.523**	0.486	0.49	0.494
200	4	1	**81.90%**	73.70%	73.70%	73.80%	**0.565**	0.513	0.515	0.517
200	5	1	**79.70%**	65.00%	65.10%	65.20%	**0.564**	0.43	0.431	0.432
200	6	1	**80.60%**	69.60%	69.70%	70.00%	**0.581**	0.483	0.486	0.489
200	7	1	**76.60%**	64.50%	64.60%	64.80%	**0.51**	0.421	0.421	0.423
200	8	1	**78.20%**	60.40%	60.40%	60.80%	**0.555**	0.368	0.367	0.371
500	1	1	**76.30%**	64.90%	65.00%	65.10%	**0.495**	0.43	0.431	0.433
500	2	1	**76.00%**	71.70%	71.80%	71.90%	0.484	0.496	0.497	**0.498**
500	3	1	**76.90%**	69.00%	69.10%	69.40%	**0.502**	0.474	0.476	0.48
500	4	1	**80.10%**	73.00%	73.00%	72.90%	**0.535**	0.5	0.5	0.499
500	5	1	**78.30%**	65.10%	65.20%	65.30%	**0.549**	0.436	0.437	0.439
500	6	1	**80.00%**	69.10%	69.30%	69.50%	**0.58**	0.481	0.483	0.487
500	7	1	**75.10%**	64.50%	64.50%	64.50%	**0.486**	0.423	0.423	0.423
500	8	1	**76.80%**	59.90%	59.90%	60.10%	**0.533**	0.362	0.363	0.364

When a separate data set was used for classification, Model 1 had the highest average classification rate across repetitions, except for the one case where the sample size was small and the students were mostly high ability students. In this case, Model 2 had a slightly better classification rate (see Table 36). The adjusted Rand index, however, did not always follow the classification rate pattern and, particularly when the sample size was small, indicated that Model 2 had the higher agreement.

Table 36: Classification information when Model 2 generated the data and a separate data set was used for classification

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
50	1	1	**67.60%**	63.30%	62.90%	62.40%	0.363	**0.412**	0.407	0.405
50	2	1	68.20%	**69.80%**	69.50%	69.70%	0.366	0.47	0.47	**0.473**
50	3	1	**69.10%**	66.30%	65.50%	64.90%	0.382	**0.437**	0.427	0.426
50	4	1	**74.80%**	71.90%	71.40%	70.30%	0.446	**0.479**	0.476	0.469
50	5	1	**70.90%**	62.90%	62.80%	62.30%	**0.432**	0.404	0.401	0.4
50	6	1	**69.80%**	67.20%	67.00%	66.40%	0.412	**0.442**	0.44	0.434
50	7	1	**66.00%**	62.90%	62.50%	62.50%	0.356	**0.415**	0.408	0.412
50	8	1	**66.70%**	55.60%	55.50%	54.70%	**0.395**	0.3	0.304	0.293
200	1	1	**73.30%**	64.40%	64.40%	64.40%	**0.449**	0.429	0.43	0.431
200	2	1	**72.60%**	70.70%	70.40%	70.20%	0.43	**0.48**	0.477	0.475
200	3	1	**73.80%**	68.50%	68.30%	68.00%	0.458	**0.465**	0.464	0.463
200	4	1	**76.80%**	71.80%	71.60%	70.90%	0.48	**0.484**	0.483	0.479
200	5	1	**75.00%**	64.60%	64.60%	64.30%	**0.498**	0.433	0.432	0.43
200	6	1	**76.10%**	68.10%	67.80%	67.70%	**0.52**	0.462	0.459	0.458
200	7	1	**71.60%**	64.40%	64.20%	64.00%	**0.431**	0.42	0.417	0.415
200	8	1	**73.30%**	59.20%	59.10%	58.70%	**0.48**	0.351	0.351	0.347
500	1	1	**74.60%**	64.20%	64.10%	64.20%	**0.471**	0.426	0.426	0.428
500	2	1	**74.00%**	71.10%	70.90%	70.80%	0.453	**0.488**	0.487	0.486
500	3	1	**75.00%**	68.40%	68.30%	68.20%	**0.477**	0.47	0.469	0.47
500	4	1	**78.30%**	72.90%	72.80%	72.50%	**0.51**	0.504	0.503	0.501
500	5	1	**76.20%**	64.50%	64.50%	64.60%	**0.516**	0.432	0.432	0.434
500	6	1	**77.80%**	68.30%	68.30%	68.20%	**0.544**	0.465	0.466	0.467
500	7	1	**72.90%**	64.40%	64.40%	64.40%	**0.451**	0.421	0.421	0.421
500	8	1	**74.80%**	59.60%	59.50%	59.40%	**0.502**	0.361	0.361	0.359

Model 1 had the highest classification rates and the highest adjusted Rand index when Model 3 was the generating model and the same data was used for classification as for generating the parameters (see Table 37).

Table 37: Classification information when Model 3 generated the data and the same data was used for classification.

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
50	1	1	**86.60%**	73.60%	79.70%	78.30%	**0.671**	0.503	0.559	0.549
50	2	1	**83.50%**	66.70%	68.80%	68.80%	**0.612**	0.464	0.483	0.484
50	3	1	**87.60%**	75.00%	82.20%	81.20%	**0.694**	0.557	0.611	0.605
50	4	1	**85.50%**	75.30%	78.90%	79.10%	**0.653**	0.569	0.589	0.593
50	5	1	**86.40%**	75.90%	81.90%	80.40%	**0.67**	0.555	0.6	0.596
50	6	1	**84.40%**	71.40%	74.30%	74.00%	**0.626**	0.514	0.542	0.538
50	7	1	**86.30%**	80.00%	82.10%	81.00%	**0.646**	0.562	0.577	0.568
50	8	1	**86.70%**	74.10%	75.00%	74.70%	**0.661**	0.532	0.544	0.545
50	9	1	**85.70%**	67.50%	69.20%	69.40%	**0.66**	0.476	0.493	0.501
50	10	1	**86.40%**	71.30%	72.30%	72.00%	**0.675**	0.503	0.515	0.518
50	11	1	**86.10%**	70.60%	73.50%	73.70%	**0.676**	0.515	0.555	0.561
50	12	1	**82.10%**	65.30%	67.10%	68.00%	**0.595**	0.439	0.455	0.467
50	13	1	**83.20%**	66.20%	69.40%	69.80%	**0.621**	0.461	0.493	0.497
50	14	1	**82.60%**	65.70%	67.40%	68.00%	**0.606**	0.444	0.463	0.471
50	15	1	**84.30%**	61.90%	67.80%	68.60%	**0.643**	0.419	0.491	0.505
50	16	1	**84.40%**	61.90%	62.60%	64.40%	**0.649**	0.381	0.395	0.412
200	1	1	**82.30%**	71.20%	78.90%	78.30%	**0.588**	0.472	0.532	0.528
200	2	1	**78.70%**	65.20%	66.40%	66.60%	**0.53**	0.441	0.451	0.454
200	3	1	**83.50%**	72.80%	81.80%	81.60%	**0.619**	0.536	0.602	0.601
200	4	1	**80.30%**	72.40%	76.60%	76.70%	0.562	0.527	0.564	**0.566**
200	5	1	**82.40%**	73.00%	79.80%	79.40%	**0.591**	0.511	0.553	0.551
200	6	1	**79.30%**	68.50%	70.60%	71.00%	**0.54**	0.478	0.5	0.505
200	7	1	**83.10%**	78.70%	81.40%	81.20%	**0.583**	0.538	0.556	0.557
200	8	1	**82.20%**	71.60%	72.10%	72.30%	**0.573**	0.499	0.499	0.504
200	9	1	**79.50%**	65.40%	66.60%	67.00%	**0.559**	0.451	0.463	0.467
200	10	1	**80.70%**	69.00%	69.40%	69.70%	**0.575**	0.468	0.473	0.476
200	11	1	**81.40%**	68.70%	70.20%	70.30%	**0.6**	0.496	0.517	0.521
200	12	1	**77.30%**	64.00%	64.50%	64.90%	**0.515**	0.423	0.428	0.431
200	13	1	**78.40%**	64.80%	67.00%	67.20%	**0.538**	0.44	0.464	0.466
200	14	1	**77.10%**	65.10%	65.70%	66.00%	**0.51**	0.433	0.438	0.441
200	15	1	**79.50%**	60.80%	64.80%	65.00%	**0.577**	0.394	0.456	0.458
200	16	1	**77.90%**	59.40%	59.50%	59.80%	**0.55**	0.357	0.357	0.362

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
500	1	1	**81.20%**	70.70%	78.50%	78.30%	**0.571**	0.463	0.523	0.522
500	2	1	**78.40%**	65.30%	66.20%	66.60%	**0.529**	0.448	0.456	0.459
500	3	1	**82.30%**	72.30%	81.40%	81.30%	**0.605**	0.53	0.595	0.595
500	4	1	**79.40%**	72.70%	76.40%	76.50%	0.554	0.537	**0.567**	0.567
500	5	1	**81.20%**	72.80%	79.50%	79.20%	**0.571**	0.509	0.544	0.543
500	6	1	**78.40%**	67.60%	69.70%	69.80%	**0.527**	0.469	0.49	0.491
500	7	1	**82.30%**	78.60%	81.20%	81.10%	**0.568**	0.533	0.55	0.55
500	8	1	**81.00%**	71.00%	71.50%	71.60%	**0.551**	0.485	0.484	0.487
500	9	1	**78.50%**	65.00%	65.80%	65.90%	**0.543**	0.448	0.456	0.457
500	10	1	**79.50%**	68.30%	68.80%	69.00%	**0.562**	0.468	0.475	0.477
500	11	1	**79.70%**	68.20%	69.20%	69.30%	**0.571**	0.489	0.5	0.504
500	12	1	**76.10%**	64.60%	64.90%	65.00%	**0.495**	0.428	0.43	0.431
500	13	1	**76.90%**	64.60%	66.00%	66.20%	**0.512**	0.44	0.457	0.458
500	14	1	**75.80%**	64.00%	64.20%	64.30%	**0.492**	0.421	0.423	0.423
500	15	1	**78.60%**	60.00%	64.00%	64.20%	**0.563**	0.385	0.449	0.451
500	16	1	**76.90%**	60.10%	60.20%	60.30%	**0.535**	0.366	0.366	0.368

When a separate sample was used and the sample size was large, then Model 1 had the highest classification rate (although there were two cases where Model 3 had the same classification rate) (see

Table 38 38). However, when the sample size was small, Model 3 had similar or higher classification rates when there was a strict hierarchy, while Model 1 performed better when there was not a strict hierarchy. The adjusted Rand index provided similar results, although when the sample size was small and the model would accept attributes of Levels 2 or 3 in either order, this index was highest for Model 3.

Table 38: Classification information when Model 3 generated the data and a separate data set was used for classification.

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
50	1	1	73.50%	70.60%	**75.40%**	74.10%	0.442	0.459	**0.486**	0.479
50	2	1	**69.80%**	62.60%	63.00%	63.00%	0.397	0.415	0.415	**0.417**
50	3	1	74.20%	70.00%	**78.20%**	77.00%	0.489	0.487	**0.566**	0.563
50	4	1	72.50%	70.20%	**74.20%**	73.70%	0.452	0.505	**0.539**	0.539
50	5	1	72.50%	69.60%	**75.40%**	73.90%	0.433	0.451	**0.494**	0.486
50	6	1	**69.90%**	65.90%	66.70%	67.00%	0.396	0.442	0.45	**0.462**
50	7	1	75.10%	76.40%	**78.50%**	77.50%	0.437	0.498	**0.515**	0.506
50	8	1	**73.70%**	68.40%	68.60%	68.50%	0.422	0.454	0.454	**0.455**
50	9	1	**71.20%**	62.70%	63.20%	62.70%	**0.444**	0.421	0.429	0.429
50	10	1	**71.30%**	64.90%	65.30%	65.20%	**0.437**	0.427	0.433	0.433
50	11	1	**72.60%**	64.80%	66.20%	65.80%	**0.468**	0.442	0.457	0.455
50	12	1	**66.70%**	62.20%	62.60%	62.40%	0.354	0.397	**0.399**	0.397
50	13	1	**67.70%**	62.00%	63.20%	62.80%	0.38	0.41	**0.415**	0.412
50	14	1	**65.40%**	62.30%	62.30%	61.90%	0.341	**0.396**	0.39	0.389
50	15	1	**67.80%**	57.40%	60.80%	60.40%	**0.415**	0.359	0.4	0.4
50	16	1	**67.30%**	57.40%	56.90%	56.80%	**0.398**	0.328	0.328	0.326
200	1	1	**78.50%**	70.00%	77.00%	76.50%	**0.527**	0.454	0.503	0.5
200	2	1	**74.70%**	64.10%	64.30%	64.40%	**0.468**	0.428	0.429	0.43
200	3	1	79.90%	71.40%	**80.60%**	80.30%	0.572	0.516	**0.588**	0.587
200	4	1	**76.60%**	71.90%	75.50%	75.40%	0.51	0.526	**0.555**	0.555
200	5	1	**78.10%**	71.60%	78.10%	77.70%	0.518	0.482	**0.522**	0.521
200	6	1	**75.10%**	66.70%	68.70%	68.40%	**0.475**	0.455	0.472	0.472
200	7	1	80.00%	78.00%	**80.50%**	80.10%	0.534	0.527	**0.543**	0.543
200	8	1	**78.00%**	69.50%	69.50%	69.40%	**0.505**	0.466	0.465	0.466
200	9	1	**75.10%**	63.80%	64.00%	64.00%	**0.493**	0.432	0.434	0.436
200	10	1	**76.90%**	67.90%	68.10%	67.80%	**0.517**	0.46	0.463	0.463
200	11	1	**77.00%**	67.20%	67.50%	67.60%	**0.532**	0.477	0.481	0.486
200	12	1	**71.90%**	64.50%	64.50%	64.30%	**0.425**	0.425	0.425	0.424
200	13	1	**73.40%**	63.90%	64.70%	64.50%	**0.455**	0.426	0.435	0.433
200	14	1	**72.40%**	64.40%	64.40%	64.20%	**0.438**	0.425	0.425	0.424
200	15	1	**75.30%**	59.70%	62.70%	62.40%	**0.514**	0.383	0.425	0.423
200	16	1	**73.30%**	59.20%	59.10%	58.80%	**0.479**	0.354	0.354	0.35

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
500	1	1	**80.10%**	70.30%	78.10%	77.80%	**0.556**	0.461	0.52	0.519
500	2	1	**76.30%**	64.60%	64.80%	64.70%	**0.493**	0.434	0.435	0.436
500	3	1	**81.10%**	72.00%	**81.10%**	81.00%	0.595	0.53	**0.596**	0.596
500	4	1	**77.80%**	72.40%	76.10%	76.00%	0.532	0.532	**0.564**	0.563
500	5	1	**78.80%**	72.00%	78.10%	77.90%	**0.533**	0.494	0.523	0.523
500	6	1	**76.60%**	67.10%	69.10%	68.80%	**0.502**	0.464	0.483	0.482
500	7	1	**80.50%**	78.10%	**80.50%**	80.40%	**0.541**	0.524	0.539	0.54
500	8	1	**79.00%**	69.30%	69.90%	69.80%	**0.517**	0.462	0.462	0.463
500	9	1	**76.10%**	64.50%	64.70%	64.70%	**0.506**	0.443	0.443	0.443
500	10	1	**77.80%**	68.00%	68.40%	68.30%	**0.534**	0.466	0.473	0.473
500	11	1	**77.60%**	67.30%	67.90%	67.80%	**0.538**	0.479	0.485	0.487
500	12	1	**73.90%**	64.10%	64.20%	64.20%	**0.46**	0.421	0.423	0.423
500	13	1	**75.40%**	64.30%	65.20%	65.20%	**0.486**	0.437	0.447	0.447
500	14	1	**73.80%**	64.00%	64.10%	64.00%	**0.458**	0.418	0.419	0.419
500	15	1	**77.50%**	60.40%	63.20%	63.10%	**0.548**	0.387	0.436	0.433
500	16	1	**75.30%**	60.10%	60.00%	59.80%	**0.511**	0.367	0.367	0.365

When Model 4 was the generating model and the same data was used for classification as for generating the data, then again Model 1 had the highest classification rates. The adjusted Rand index also was highest for Model 1 except in the large sample size when there were mostly high ability students, and something other than a strict hierarchy was used. In these cases the adjusted Rand index indicated Models 3 or 4 (see Table 39).

Table 39: Classification information when Model 4 generated the data and the same data was used for classification

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
50	1	1	**86.60%**	73.10%	79.70%	78.70%	**0.674**	0.502	0.558	0.555
50	2	1	**84.10%**	67.20%	69.00%	69.10%	**0.624**	0.467	0.485	0.487
50	3	1	**84.60%**	67.30%	68.70%	69.50%	**0.638**	0.458	0.472	0.483
50	4	1	**83.80%**	68.50%	70.00%	70.30%	**0.621**	0.487	0.503	0.509
50	5	1	**87.50%**	73.10%	82.30%	80.50%	**0.702**	0.535	0.617	0.604
50	6	1	**85.10%**	73.50%	77.00%	76.90%	**0.646**	0.54	0.572	0.573
50	7	1	**84.10%**	73.40%	74.40%	74.90%	**0.625**	0.53	0.536	0.536
50	8	1	**84.50%**	73.20%	76.20%	76.10%	**0.633**	0.529	0.558	0.56
50	9	1	**86.60%**	74.80%	80.30%	78.80%	**0.669**	0.518	0.563	0.558
50	10	1	**84.00%**	71.50%	74.00%	73.90%	**0.622**	0.513	0.537	0.54
50	11	1	**84.20%**	71.80%	73.40%	73.30%	**0.633**	0.523	0.539	0.543
50	12	1	**84.30%**	72.60%	74.70%	74.50%	**0.627**	0.531	0.549	0.553
50	13	1	**87.10%**	81.10%	83.50%	82.10%	**0.659**	0.57	0.591	0.577
50	14	1	**86.70%**	76.30%	76.90%	76.70%	**0.66**	0.553	0.562	0.562
50	15	1	**86.40%**	75.90%	77.00%	76.30%	**0.649**	0.541	0.56	0.554
50	16	1	**86.70%**	75.90%	76.50%	75.70%	**0.662**	0.55	0.556	0.559
50	17	1	**84.60%**	66.20%	66.50%	67.90%	**0.631**	0.441	0.442	0.461
50	18	1	**85.20%**	68.20%	70.90%	71.20%	**0.65**	0.492	0.526	0.534
50	19	1	**85.90%**	72.20%	72.60%	73.30%	**0.66**	0.519	0.526	0.534
50	20	1	**86.50%**	70.70%	74.00%	73.40%	**0.686**	0.516	0.56	0.556
50	21	1	**84.20%**	66.00%	67.40%	68.90%	**0.641**	0.443	0.455	0.472
50	22	1	**83.70%**	65.40%	69.10%	69.10%	**0.631**	0.438	0.481	0.479
50	23	1	**86.60%**	62.50%	64.90%	67.80%	**0.712**	0.418	0.459	0.491
50	24	1	**84.50%**	62.90%	66.90%	68.00%	**0.653**	0.428	0.49	0.499

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
200	1	1	**82.60%**	71.60%	78.40%	77.90%	**0.596**	0.476	0.526	0.524
200	2	1	**79.20%**	65.70%	67.10%	67.30%	**0.537**	0.441	0.456	0.458
200	3	1	**78.90%**	65.50%	66.30%	67.10%	**0.533**	0.442	0.449	0.456
200	4	1	**79.50%**	65.30%	66.60%	66.90%	**0.545**	0.445	0.456	0.459
200	5	1	**82.60%**	71.30%	80.40%	80.20%	**0.607**	0.502	0.58	0.578
200	6	1	**79.70%**	71.60%	74.40%	74.40%	**0.55**	0.519	0.544	0.545
200	7	1	**78.30%**	71.20%	72.20%	72.90%	**0.521**	0.507	0.516	0.519
200	8	1	**79.40%**	71.40%	74.50%	74.50%	**0.546**	0.515	0.541	0.542
200	9	1	**81.70%**	72.70%	79.20%	78.80%	**0.576**	0.503	0.541	0.539
200	10	1	**79.40%**	69.30%	71.40%	71.30%	**0.544**	0.492	0.511	0.51
200	11	1	**79.00%**	69.40%	70.90%	71.20%	**0.537**	0.488	0.506	0.51
200	12	1	**79.70%**	69.60%	71.70%	71.80%	**0.547**	0.49	0.509	0.513
200	13	1	**83.30%**	78.60%	81.30%	80.90%	**0.585**	0.536	0.553	0.551
200	14	1	**82.10%**	73.60%	74.00%	73.90%	**0.57**	0.512	0.517	0.518
200	15	1	**81.90%**	73.20%	73.70%	73.60%	**0.568**	0.505	0.514	0.514
200	16	1	**82.10%**	74.00%	74.50%	74.20%	**0.573**	0.519	0.523	0.524
200	17	1	**81.20%**	66.00%	66.40%	67.30%	**0.588**	0.446	0.449	0.459
200	18	1	**79.80%**	65.30%	66.70%	66.80%	**0.566**	0.449	0.463	0.463
200	19	1	**82.10%**	69.30%	70.10%	71.40%	**0.608**	0.492	0.498	0.514
200	20	1	**81.60%**	69.70%	71.20%	71.20%	**0.604**	0.511	0.53	0.534
200	21	1	**78.10%**	64.50%	65.20%	66.50%	**0.531**	0.422	0.429	0.443
200	22	1	**77.80%**	64.70%	66.70%	67.20%	**0.524**	0.444	0.464	0.468
200	23	1	**80.60%**	60.40%	62.90%	65.30%	**0.604**	0.385	0.433	0.465
200	24	1	**79.70%**	60.20%	64.40%	64.90%	**0.578**	0.387	0.451	0.457

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
500	1	1	**80.70%**	70.70%	78.00%	77.70%	**0.562**	0.46	0.514	0.513
500	2	1	**78.00%**	65.00%	65.50%	65.90%	**0.518**	0.436	0.441	0.444
500	3	1	**77.90%**	64.90%	65.30%	65.90%	**0.518**	0.437	0.441	0.447
500	4	1	**77.50%**	64.40%	65.10%	65.30%	**0.508**	0.428	0.433	0.436
500	5	1	**81.70%**	70.80%	80.40%	80.30%	**0.591**	0.5	0.574	0.574
500	6	1	**78.20%**	70.80%	73.90%	74.00%	0.524	0.509	0.537	**0.539**
500	7	1	**76.20%**	70.60%	71.40%	72.10%	0.487	0.499	0.506	**0.507**
500	8	1	**78.30%**	70.90%	73.80%	73.90%	0.527	0.511	**0.538**	0.538
500	9	1	**80.90%**	72.30%	79.00%	78.80%	**0.564**	0.501	0.536	0.535
500	10	1	**78.60%**	69.10%	71.10%	71.30%	**0.531**	0.487	0.506	0.508
500	11	1	**77.30%**	68.30%	69.70%	70.00%	**0.507**	0.476	0.492	0.495
500	12	1	**78.10%**	68.40%	70.60%	70.80%	**0.521**	0.477	0.496	0.498
500	13	1	**82.20%**	78.40%	81.10%	80.90%	**0.565**	0.524	0.545	0.544
500	14	1	**81.00%**	73.60%	73.70%	73.70%	**0.552**	0.513	0.512	0.512
500	15	1	**80.80%**	73.30%	73.30%	73.40%	**0.549**	0.511	0.511	0.512
500	16	1	**80.80%**	73.60%	73.60%	73.70%	**0.549**	0.514	0.511	0.513
500	17	1	**79.60%**	64.80%	65.00%	65.80%	**0.563**	0.432	0.433	0.44
500	18	1	**78.20%**	64.80%	65.30%	65.60%	**0.536**	0.444	0.452	0.456
500	19	1	**80.50%**	68.80%	69.20%	70.20%	**0.579**	0.482	0.485	0.496
500	20	1	**79.90%**	68.80%	69.80%	70.00%	**0.571**	0.497	0.51	0.513
500	21	1	**76.70%**	64.50%	64.80%	66.30%	**0.506**	0.428	0.431	0.447
500	22	1	**77.00%**	64.80%	66.30%	66.60%	**0.513**	0.442	0.461	0.463
500	23	1	**79.10%**	60.20%	62.20%	64.60%	**0.585**	0.38	0.43	0.462
500	24	1	**78.50%**	59.90%	63.60%	64.10%	**0.564**	0.384	0.445	0.451

When Model 4 was the generating model and a separate data set was used for classification then for large sample sizes, Model 1 seemed to have the highest classification rate (see Table 40). When a sample size of 200 was used, Model 1 had the highest classification rate except for when the students' ability distribution was skewed and a strict hierarchy was followed. The small sample size also followed this trend, although Model 3 had a higher classification rate for a couple more cells in which a strict hierarchy was followed. The adjusted Rand index seemed to indicate Models 3 or 4 more often when the sample size was low, but tended to agree with the classification rates when the sample size was high.

Table 40: Classification information when Model 4 generated the data and a separate data set was used for classification

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
50	1	1	73.00%	69.20%	**75.20%**	73.80%	0.438	0.444	**0.48**	0.472
50	2	1	**69.30%**	63.80%	64.00%	63.90%	0.38	0.419	**0.423**	0.423
50	3	1	**69.10%**	63.60%	63.20%	62.90%	0.376	**0.406**	0.4	0.4
50	4	1	**69.70%**	63.60%	64.70%	64.40%	0.387	0.418	0.425	**0.426**
50	5	1	72.00%	67.90%	**76.50%**	74.90%	0.438	0.453	**0.517**	0.513
50	6	1	69.30%	67.80%	**70.70%**	70.10%	0.386	0.453	**0.478**	0.476
50	7	1	**69.00%**	67.70%	68.50%	68.60%	0.371	0.457	**0.463**	0.461
50	8	1	69.30%	67.30%	69.90%	**70.10%**	0.396	0.454	0.477	**0.483**
50	9	1	72.30%	69.40%	**74.90%**	73.20%	0.423	0.452	**0.486**	0.48
50	10	1	**70.30%**	66.90%	69.00%	68.50%	0.406	0.449	0.471	**0.473**
50	11	1	**70.70%**	67.40%	68.30%	67.50%	0.412	0.469	**0.474**	0.469
50	12	1	**69.10%**	65.80%	67.40%	66.50%	0.391	0.444	**0.456**	0.452
50	13	1	75.00%	76.10%	**78.00%**	77.00%	0.432	0.49	**0.503**	0.497
50	14	1	**74.00%**	70.10%	69.80%	69.00%	0.428	0.453	**0.453**	0.444
50	15	1	**74.80%**	71.10%	70.60%	69.80%	0.45	**0.473**	0.47	0.469
50	16	1	**74.20%**	69.90%	70.10%	68.60%	0.428	0.45	**0.456**	0.441
50	17	1	**70.50%**	61.70%	61.10%	61.80%	**0.424**	0.388	0.383	0.39
50	18	1	**70.20%**	63.20%	63.40%	63.10%	0.431	0.433	**0.435**	0.434
50	19	1	**73.50%**	66.70%	66.70%	67.30%	**0.468**	0.463	0.457	0.463
50	20	1	**73.50%**	67.60%	68.10%	67.00%	0.474	0.473	**0.476**	0.474
50	21	1	**68.10%**	61.90%	62.60%	63.10%	0.372	0.393	0.397	**0.405**
50	22	1	**67.30%**	63.30%	64.10%	63.60%	0.371	0.414	**0.425**	0.423
50	23	1	**69.50%**	57.60%	60.10%	61.80%	**0.453**	0.365	0.403	0.427
50	24	1	**69.00%**	57.20%	60.80%	61.00%	**0.427**	0.355	0.403	0.409

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
200	1	1	**78.00%**	69.50%	76.50%	75.90%	**0.517**	0.445	0.495	0.492
200	2	1	**74.50%**	63.80%	64.30%	64.30%	**0.458**	0.42	0.424	0.424
200	3	1	**74.90%**	64.20%	64.30%	64.60%	**0.472**	0.424	0.427	0.431
200	4	1	**74.70%**	64.00%	64.20%	64.20%	**0.467**	0.422	0.423	0.427
200	5	1	79.00%	69.60%	**79.50%**	79.40%	0.55	0.48	0.564	**0.567**
200	6	1	**75.30%**	70.00%	72.90%	73.10%	0.476	0.49	0.521	**0.524**
200	7	1	**73.50%**	69.30%	70.20%	70.60%	0.446	0.481	0.491	**0.492**
200	8	1	**75.40%**	69.90%	73.00%	72.90%	0.482	0.493	**0.524**	0.523
200	9	1	78.40%	72.00%	**78.50%**	78.10%	0.522	0.493	**0.533**	0.53
200	10	1	**75.40%**	68.00%	69.90%	69.70%	0.476	0.468	**0.483**	0.483
200	11	1	**75.20%**	68.60%	69.40%	69.30%	0.479	0.482	**0.493**	0.493
200	12	1	**75.30%**	68.80%	70.20%	69.80%	0.482	0.481	**0.492**	0.492
200	13	1	79.60%	77.30%	**80.60%**	80.30%	0.526	0.517	**0.543**	0.541
200	14	1	**78.60%**	72.70%	73.00%	72.50%	**0.515**	0.502	0.503	0.501
200	15	1	**77.80%**	72.00%	71.90%	71.40%	**0.495**	0.487	0.485	0.483
200	16	1	**77.90%**	72.30%	72.40%	72.00%	**0.504**	0.495	0.497	0.495
200	17	1	**76.70%**	64.90%	64.80%	65.40%	**0.521**	0.436	0.437	0.444
200	18	1	**75.40%**	64.00%	64.50%	64.40%	**0.501**	0.433	0.438	0.44
200	19	1	**76.70%**	67.80%	68.00%	68.20%	**0.517**	0.471	0.473	0.475
200	20	1	**76.30%**	68.60%	68.20%	68.10%	**0.519**	0.495	0.489	0.492
200	21	1	**73.80%**	64.20%	64.20%	65.10%	**0.459**	0.421	0.422	0.431
200	22	1	**74.10%**	64.30%	65.30%	65.10%	**0.466**	0.436	0.447	0.446
200	23	1	**75.30%**	59.30%	61.10%	62.90%	**0.527**	0.372	0.41	0.44
200	24	1	**75.40%**	58.90%	62.60%	62.80%	**0.516**	0.373	0.43	0.432

SS	LP	OV	Percent of students correctly classified				adjusted Rand Index			
			M1	M2	M3	M4	M1	M2	M3	M4
500	1	1	**79.70%**	70.00%	77.70%	77.50%	**0.548**	0.454	0.513	0.512
500	2	1	**76.70%**	64.30%	64.40%	64.50%	**0.498**	0.426	0.427	0.428
500	3	1	**75.80%**	64.40%	64.60%	64.80%	**0.484**	0.429	0.431	0.434
500	4	1	**76.10%**	64.00%	64.30%	64.40%	**0.489**	0.426	0.429	0.431
500	5	1	**80.40%**	69.70%	80.00%	79.80%	**0.574**	0.485	0.57	0.57
500	6	1	**76.80%**	70.20%	73.00%	73.10%	0.506	0.5	0.528	**0.529**
500	7	1	**74.50%**	70.20%	70.80%	71.30%	0.461	0.493	**0.5**	0.5
500	8	1	**76.60%**	70.60%	73.40%	73.50%	0.503	0.506	0.533	**0.534**
500	9	1	**79.60%**	72.40%	78.80%	78.50%	**0.549**	0.505	0.536	0.536
500	10	1	**76.60%**	68.50%	70.10%	69.90%	**0.501**	0.479	0.492	0.492
500	11	1	**75.90%**	68.50%	69.50%	69.40%	0.491	0.483	0.492	**0.493**
500	12	1	**76.50%**	68.40%	70.10%	69.80%	**0.5**	0.48	0.494	0.493
500	13	1	80.40%	78.00%	**80.50%**	80.40%	**0.539**	0.522	0.538	0.538
500	14	1	**78.90%**	72.70%	72.70%	72.70%	**0.518**	0.497	0.496	0.497
500	15	1	**79.00%**	72.90%	72.80%	72.70%	**0.521**	0.506	0.505	0.505
500	16	1	**79.10%**	73.00%	73.00%	72.90%	**0.523**	0.503	0.502	0.503
500	17	1	**77.70%**	64.30%	64.30%	64.70%	**0.534**	0.426	0.427	0.431
500	18	1	**76.00%**	64.10%	64.30%	64.30%	**0.505**	0.434	0.435	0.437
500	19	1	**78.50%**	68.40%	68.70%	69.30%	**0.551**	0.479	0.481	0.489
500	20	1	**77.70%**	68.20%	68.60%	68.50%	**0.538**	0.491	0.494	0.496
500	21	1	**75.10%**	64.40%	64.50%	65.60%	**0.48**	0.426	0.427	0.438
500	22	1	**75.30%**	64.40%	65.30%	65.20%	**0.484**	0.438	0.448	0.447
500	23	1	**76.70%**	59.90%	61.70%	63.50%	**0.545**	0.377	0.419	0.444
500	24	1	**77.30%**	59.50%	62.70%	62.80%	**0.549**	0.38	0.436	0.434

Discussion

Under all the conditions for which the data was generated, Model 1 performed well both when it was the correct model and when it was incorrectly specified. Parameters for this model were recovered a high percentage of the time, often this was the best fitting model and this model often had the highest classification rates, particularly when the sample size was large and/or the students were equally distributed along the ability spectrum. Even when Model 1 did not have the highest classification rate, it still had classification rates that were fairly close to the other models.

Model 3 also performs well, particularly when the sample size is small or the ability distribution of the students is skewed. While Models 2 and 4 also had high parameter recovery rates and high classification rates, their classification rates were not generally quite as high as Model 1 or Model 3 and the fit statistics did not often pick these models as the best fitting models.

Conclusion

Overall, a practitioner would want to use Model 1 or Model 3. When making the choice between these models, a practitioner should take into account the theoretical background of the learning progression as well as the target subject. Their decision should be influenced by their belief on the true underlying structure of the levels of the learning progression and the relationships between different attributes. Practitioners should also take into account the interpretation of the attributes and how that would affect the students. When using Model 1 when the true model is not strictly linear, students may be classified in a low level, even if they have some of the attributes at the higher level. It is important to consider how these misclassifications would affect the student and the importance of being able to distinguish students who follow different learning paths.

In addition, this research indicated that Model 3 might provide more accurate classifications when the ability levels of the student was skewed, or when the sample size was small. Practitioners who are working in these types of environments may want to consider the tradeoffs involved in using a model with more variables (Model 3) versus a model that may not classify as many students correctly (Model 1). In general, the recommendation is to use Model 1, even in cases where a strict hierarchy might not be followed and where students are allowed to follow multiple paths.

CHAPTER 5: MODELING TWO LEARNING PROGRESSIONS

This second study will focus on issues surrounding the use of two learning progression variables in an assessment, particularly issues surrounding the structure of observable variables' dependence on latent variable student-model parents. While there may be hypotheses regarding how the influence of multiple learning progression variables on task performance can combine (based on the underlying substantive theory of the learning progressions), the structure of this relationship may not be known in advance in a real data situation. This study will address a question of robustness; whether there are certain situations in which a more constrained or less constrained model would provide comparable or more accurate results than the generating model, from among a set of paradigmatic model structures.

Study Overview

This study will focus on the conditional probabilities of observables variables given values of (latent) proficiency variables that reflect learning progressions. Note that while the first study contained models that had different graphical structures, in this study the nature of the learning progression will be the same across all of the different models. They will all follow Figure 18, and have one latent variable representing each LP and all items depending on both of these LPs. Instead this study examines a second question which may come up when using a BIN, which involves examining benefits and/or drawbacks to placing constraints on the conditional probabilities for observable variables given the LP variables. It will compare the unconstrained estimation of these conditional probabilities (i.e., a hyper-Dirichlet conditional probability matrix) and the compensatory, conjunctive and disjunctive models (see Table 41). The research sub-questions here will be the same as in study 1:

1) How well are parameters recovered under each model for the various conditions?

2) How do inferences regarding students (i.e., posterior distributions for proficiency variable) compare across the different models under various conditions?

3) How do goodness-of-fit tests perform at identifying the correct model under various conditions?

Table 41: Probability constraints for the different models. J is the vector containing the level on LP1 and LP2

Model	Probability constraints
1: Unconstrained	$P\ (X_{i\bar{J}} = 1) = P(X_i \mid \bar{J})$
2: Compensatory	$P(X_i \mid \bar{J}) = \dfrac{\exp(\sum_J \xi_j + \sigma_i)}{1 + \exp(\sum_J \xi_j + \sigma_i)}$
3: Conjunctive	$P(X_i \mid \bar{J}) = \dfrac{\exp(\min(\bar{\xi}) + \sigma_i)}{1 + \exp(\min(\bar{\xi}) + \sigma_i)}$
4: Disjunctive	$P(X_i \mid \bar{J}) = \dfrac{\exp(\max(\bar{\xi}) + \sigma_i)}{1 + \exp(\max(\bar{\xi}) + \sigma_i)}$

These questions will again be addressed by a simulation. Data will be simulated based on each of the different models, then estimation results will be computed for each of the given models and the results will be compared. The overall conditional probabilities will be compared, along with the categorization of each student on both of the learning progressions. The simulation will again examine how sample size, number of observables and different parameter structures affect the different models.

Study Conditions

This study will again use four levels for each learning progression, with a novice level for students who do not any attributes (for a total of five levels). As noted before, four levels is a common number of levels used in practice, and four levels provide enough flexibility to allow for differentiation between the models without overly complicating the situation.

This study will examine the case where the probability of membership for each of the levels of the learning progression is equal (or .2). In condition 1 the two learning progressions will be independent of each other, marking the case where students may be at any combination of levels of the two LPs. The second condition will have the two learning progressions highly related, indicating a relationship between the two learning progressions. In this case, students have high probability of being at the same level in both of the learning progressions, a lower probability of being at levels one away from each other, and an even lower probability of being at levels that are further apart (see Table 42). The Pearson correlation between these two LPs is 0.73.

Table 42: Probability that a student is at a combination of levels of each of the LPs

Level of LP2	Level of LP 1					Total
	1	2	3	4	5	
1	0.15	0.02	0.01	0.01	0.01	0.2
2	0.02	0.14	0.02	0.01	0.01	0.2
3	0.01	0.02	0.14	0.02	0.01	0.2
4	0.01	0.01	0.02	0.14	0.02	0.2
5	0.01	0.01	0.01	0.02	0.15	0.2
Total	0.2	0.2	0.2	0.2	0.2	1

Two possibilities for the number of observables will be used. In the first there are 3 observables for each combination of levels of the LPs (not counting the novice level as items are not designed to measure that level) (16 total combination of levels, for a total of 48 items). In the second there are 30 observable variables, which will have three observables for each pairing of level skills that are either the same or one off (for example, there will be an observable designed to measure level 2 of LP1 and level 3 of LP2, level 3 of LP1 and level 2 of LP2, and one to measure level 2 of LP1 and LP2, but not one to measure level 1 of LP1 and level 3 of LP2). This follows the possible situation in which the two skills are used together and it may be hard to design observables which vary drastically on the level they require of both skills. For this study each of the observable variables will be binary.

Models 2 through 4 do not have items that depend directly on given levels of the learning progression. Instead while the effective probabilities are associated with the combination of levels, the IRT structure provides a convenient lower-dimensional structure for calculating those probabilities. The effective probabilities are calculated from the difficulty of the item and the ability associated with the students' level of the LP. The ability parameter is where the association with the levels of the LP comes into play. Each level of a LP is associated with a particular ability parameter (while for an IRT model the ability parameters are on a continuum, the BIN would categorize this continuum and provide one ability parameter for each level of the LP.) The ability parameters for each of the LPs, along with the item difficulty parameter are then used to determine the probabilities for a correct response. For these models only the case where there are 30 items will be used. Additional items would allow for more observations on different IRT difficulty values, but would not necessarily change the range of values used and would not be expected to provide much more insight into the nature of the models.

For Model 1, the conditional probabilities will be estimated directly. Similarly to study 1, the generating probabilities are as follows: a probability of .8 will be used for a correct response if the student has the appropriate skill level, and a probability of .2 will be used for a correct response if they do not. Model 1 also requires the decision to be made regarding if the student has the requisite combination of skill levels. Three different conditions will be made for this decision, each of which will follow from a different dependence relationship between the observable variables and the relationships with the LPs. These conditions will follow the compensatory, conjunctive and disjunctive model. In Model 1 this will be implemented by saying a student has the appropriate skill if the sum of the levels of the LPs they have is greater than or equal to the sum of the levels required by the item. The second type will be that a student has the skills required if they are at the levels of LP required (or higher) for both LPs, while the third will only require the student to be at the level of the LP (or higher) for one of the LPs (see Table 43).

Table 43: OV probabilities for each model type for data generated with Model 1

Condition	Initial Observable Probability = .8 if: (.2 probability otherwise)
1	LP 1 level + LP 2 level >= LP 1 level req + LP 2 level req
2	LP 1 level >= LP 1 level req and LP 2 level >= LP 2 level req
3	LP 1 level >= LP 1 level req or LP 2 level >= LP 2 level req

Models 2 through 4 will each have the same specified parameters. What changes between these is how the parameters are combined to construct the probability models (as seen in Table 41). These parameters are based on the use of the LC/RM model (Formann & Kohlmann, 1998) as discussed in Chapter 3. The initial parameters that are required are the ability parameters (or theta value), which are on an IRT scale, associated with each level of the learning progression. In this method, students who are at different levels of the learning progression are thought to have different IRT ability estimates, although students at the same level of the learning progression should have the same ability estimates.

In this case only one distribution of ability parameters will be considered. This distribution will follow the use of quantiles of the normal distribution (Almond, Yan, & Hemat, 2008). The values of (-1, -.5, 0, .5 and 1) will be used for each of the levels of the LP respectively. These same values will be used for both of the learning progressions.

For the item difficulty parameters two distributions will be used. For the first distribution, numbers between -2 and 2 will be randomly generated and then ordered such that item 1 is the easiest item and item 30 is the hardest item (see Table 44). The second distribution will use values between -1.5 and 1.5 and the difficulty values will be based on the levels of the LPs that the item was designed to measure, in such a matter that items that are geared towards lower levels will be easier than items that are geared towards higher levels (see Table 44). Both of these distributions reflect the concept that items that reflect upon lower levels are easier, but the first distribution allows items to vary in their difficulty.

Table 44: The b values for items based on which levels of the LPs they depend upon

Item	LP 1 level	LP 2 level	b value for Cond. 1	b value for Cond. 2
1	1	1	-1.73	-1.5
2	1	1	-1.71	-1.5
3	1	1	-1.67	-1.5
4	1	2	-1.55	-1
5	1	2	-1.52	-1
6	1	2	-1.5	-1
7	2	1	-1.44	-1
8	2	1	-1.35	-1
9	2	1	-1.25	-1
10	2	2	-1.1	-0.05
11	2	2	-1.07	-0.05
12	2	2	-0.8	-0.05
13	2	3	-0.8	0
14	2	3	-0.71	0
15	2	3	-0.64	0
16	3	2	-0.63	0
17	3	2	-0.59	0
18	3	2	-0.27	0
19	3	3	-0.12	0.5
20	3	3	0.01	0.5
21	3	3	0.19	0.5
22	3	4	0.42	1
23	3	4	0.62	1
24	3	4	0.9	1
25	4	3	1.03	1
26	4	3	1.2	1
27	4	3	1.27	1
28	4	4	1.35	1.5
29	4	4	1.37	1.5
30	4	4	1.71	1.5

Similarly to the previous study, the sample sizes that will be used are 100 and 500. For this model, more items will be used and therefore a sample size of 50 was not deemed appropriate. A sample size of 500 was chosen to represent a large sample. In the previous study most of the insights came from the results with the small sample and the large sample, so only two samples were chosen for this study.

The total number of cells in this study was 48. Each cell was run 10 times, a number that is feasible given the long running times required in MCMC estimation. (For a sample size of 500 the minimum time it took a cell to run was 7 hours, for a sample size of 100 the minimum time was 1.5 hours. Using 10 replications and the minimum values, the total number of days the simulation would take to run is 85 days.) We will thus be able to examine main effects and qualitative differences, but not be able to estimate fine details of distributions of estimates. MCMC estimation will also be used for this study with three chains, one at the low end of the distribution, one at the middle and one at the high end of the distribution. Initial results showed that while sometimes convergence was reached with 10,000 iterations; other times more iterations were needed, therefore the study used 15,000 iterations with a burn in of 13,500. Again the Gelman-Rubin statistic was used to check for convergence.

This study also used uninformative priors in order to minimize the influence of priors on the parameter estimates. Similarly as for Study 1, Model 1 used a Dirichlet prior with a_i =2 for all the probability of being at a level of the learning progression variables and a beta distribution with a=2 and b=2 was used for the probability of the observable variables. In both of these cases, this would imply that the probability associated with the level of the learning progressions or the possible response to the items was equal for all possibilities and that the belief surrounding this was very low.

In Models 2-4 the Dirichlet prior with a_i =2 was again used for the levels of the learning progression. The item difficulty parameters associated with the learning progressions as well as the ability estimation of the students were all given priors that followed a normal distribution centered at 0 with a variance of 4. Typically IRT parameters are on a scale from -2 to 2, so having a variance of 4 is a fairly weak prior.

Model fit

This study will follow the same methods used for model fit as in Study 1. For each cell, data will be generated based on the model parameters and then all four models will be used to estimate parameters. For parameter recovery each of the models will be compared to the generating model to determine how well each model recovered parameters. The parameters that will be compared will be the probability for each level of the LPs as well as the probabilities for each observable given the different levels. For the case of Models 2-4 this will require these probabilities be computed from the ability and difficulty parameters that are recovered.

This study will again use the AIC, BIC and DIC statistics for comparing model fit. For each replication these statistics will be computed for each of the models and then the best model will be picked based on which model had the lowest value of these. Results will be compared across the statistics to determine which model seems to fit the data best.

Also included will be an examination of student-level classification; for each simulee in a given data set, the BIN built from the estimated parameters will be used to determine the most likely level for each person. This will be used to determine how well each model was able to classify the subjects. This classification will again be used on the original data used for generating the parameter estimates as well as a separate data set generated using the same parameters as the first data set. In addition, the adjusted Rand index will be used as a measure for how well the model was able to capture the correct classifications of the students.

Results

When Model 1 was the generating model, all models did a very good job at recovering the overall probability associated with the levels of the learning progression, with all cells recovering parameters over 90% of the time. Model 1 and Model 2 also had a high recovery rate for the observable variables both in the 48 observable case (see Table 45) and in the case where 30 observable variables were used (see Table 46). Models 3 and 4 were not able to recover the probabilities associated with the observable variables as well. For Models 2, 3 and 4 the recovery of the observable variable probabilities was based on using the estimates for the IRT values to calculate the conditional probabilities of responses on the observable variables. This followed from the previous study and checked to see if the 95th percentile range contained the generating value of the parameter.

Table 45: Percent of parameters recovered when Model 1 was the generating model and there were 48 observable variables

			% params rec for the LP levels				% params rec. for the OV			
	LP	OP	Model	Model	Model	Model	Model	Model	Model	Model
SS	probs	cond	1	2	3	4	1	2	3	4
100	1	1	97.0%	97.0%	98.0%	98.0%	93.7%	99.9%	70.4%	67.8%
100	1	2	96.0%	98.0%	98.0%	97.0%	94.2%	99.8%	75.4%	68.2%
100	1	3	95.0%	95.0%	96.0%	96.0%	92.6%	99.6%	64.3%	74.9%
100	2	1	97.0%	96.0%	96.0%	97.0%	92.9%	99.9%	71.6%	70.6%
100	2	2	98.0%	97.0%	97.0%	98.0%	94.7%	98.9%	77.7%	72.9%
100	2	3	98.0%	98.0%	97.0%	98.0%	92.3%	99.3%	73.9%	77.7%
500	1	1	96.0%	95.0%	96.0%	96.0%	96.2%	96.9%	57.8%	56.2%
500	1	2	96.0%	96.0%	96.0%	97.0%	96.2%	95.6%	63.6%	68.0%
500	1	3	94.0%	94.0%	94.0%	94.0%	96.2%	96.1%	70.1%	58.1%
500	2	1	94.0%	95.0%	95.0%	95.0%	95.0%	97.8%	62.9%	64.9%
500	2	2	95.0%	94.0%	96.0%	96.0%	95.1%	93.7%	62.4%	81.5%
500	2	3	90.0%	90.0%	90.0%	90.0%	94.4%	93.1%	74.8%	58.1%

Table 46: Percent of parameters recovered when Model 1 was the generating model and there were 30 observable variables

			% params rec for the LP levels				% params rec. for the OV			
	LP	OP	Model	Model	Model	Model	Model		Model	Model
SS	probs	cond	1	2	3	4	1	Model 2	3	4
100	1	1	98.0%	98.0%	98.0%	98.0%	94.00%	100.0%	86.5%	88.3%
100	1	2	94.0%	94.0%	94.0%	94.0%	94.30%	100.0%	93.3%	84.4%
100	1	3	94.0%	94.0%	94.0%	94.0%	94.20%	100.0%	84.1%	93.1%
100	2	1	95.0%	89.0%	89.0%	88.0%	94.60%	100.0%	88.5%	87.4%
100	2	2	92.0%	96.0%	95.0%	95.0%	95.40%	99.9%	92.0%	84.3%
100	2	3	97.0%	92.0%	92.0%	93.0%	94.80%	100.0%	84.4%	96.4%
500	1	1	93.0%	93.0%	94.0%	93.0%	95.70%	100.0%	74.1%	81.9%
500	1	2	96.0%	96.0%	96.0%	96.0%	96.40%	100.0%	83.3%	81.3%
500	1	3	97.0%	97.0%	97.0%	97.0%	96.30%	100.0%	94.5%	76.5%
500	2	1	93.0%	96.0%	96.0%	96.0%	95.40%	100.0%	75.9%	73.9%
500	2	2	95.0%	99.0%	99.0%	99.0%	95.20%	99.7%	88.0%	86.1%
500	2	3	96.0%	91.0%	92.0%	91.0%	95.60%	99.3%	92.9%	78.9%

One note is that when Model 1 was the generating model, Models 3 and 4 had a higher percentage recovery rate when there were 30 variables than in the case where there were 48 variables. From examining the individual cell results Models 3 and 4 varied in where they had difficulty from each. In the case where the probability was based on the summative levels of both of the LPs (case 1), Model 3 had difficulty when the item did not require one of the LP attributes (note that there were less of these items in the 30 variable case). Model 4 also had some difficulties when one of the LPs was not required, but only when the overall ability was not as high. Model 4 also had some difficulties when the overall ability was equal to or higher than the overall requirements.

In the case where the requirements of the item was based on having enough ability on both of the LPs (case 2), Model 3 had difficulties when one of the levels of the learning progression was at or one level above what the item required but the other level of the learning progression was not equal to the requirement of the item. Model 4 had difficulties when one or more of the attribute levels required were high when the item requirements were close to each other and had difficulties in recovering parameters across levels of the attributes when the requirements were further apart from each other.

In the case where the requirements of the item depended on the highest level of the LP (case 3) then both Model 3 and Model 4 had difficulty recovering the parameters when the attribute levels were low but more so when the items requirements were further apart from each other.

When Models 2 through 4 were the generating models then all models had a recovery rate of higher than 90%. The observable variable parameters were able to be recovered every time using Model 2. Models 3 and 4 also had a 100% recovery rate with their own generating OV parameters (see Table 47 - Table 49).

Table 47: Percent of parameters recovered when Model 2 was the generating model and there were 30 observable variables

SS	LP probs	OP cond	% params rec for the LP levels				% params rec. for the OV			
			Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
100	1	1	98.0%	98.0%	98.0%	98.0%	95.5%	100.0%	98.9%	99.3%
100	1	2	93.0%	94.0%	94.0%	94.0%	95.8%	100.0%	99.0%	99.4%
100	2	1	98.0%	98.0%	98.0%	98.0%	96.5%	100.0%	96.3%	96.7%
100	2	2	95.0%	95.0%	94.0%	95.0%	96.5%	100.0%	96.2%	96.5%
500	1	1	96.0%	96.0%	96.0%	97.0%	95.9%	100.0%	95.5%	98.1%
500	1	2	100.0%	100.0%	100.0%	100.0%	95.9%	100.0%	92.8%	91.6%
500	2	1	95.0%	95.0%	95.0%	95.0%	96.0%	100.0%	93.4%	94.2%
500	2	2	93.0%	94.0%	93.0%	93.0%	96.2%	100.0%	96.0%	95.0%

Table 48: Percent of parameters recovered when Model 3 was the generating model and there were 30 observable variables

SS	LP probs	OP cond	% params rec for the LP levels				% params rec. for the OV			
			Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
100	1	1	97.0%	97.0%	97.0%	97.0%	95.6%	100.0%	100.0%	99.1%
100	1	2	92.0%	94.0%	93.0%	93.0%	96.1%	100.0%	100.0%	99.2%
100	2	1	98.0%	98.0%	98.0%	98.0%	96.3%	100.0%	100.0%	96.6%
100	2	2	96.0%	96.0%	95.0%	96.0%	96.9%	100.0%	100.0%	96.3%
500	1	1	97.0%	97.0%	97.0%	97.0%	95.9%	100.0%	100.0%	98.7%
500	1	2	96.0%	96.0%	95.0%	95.0%	95.9%	100.0%	100.0%	99.3%
500	2	1	96.0%	96.0%	96.0%	96.0%	96.0%	100.0%	100.0%	94.5%
500	2	2	99.0%	99.0%	99.0%	99.0%	96.1%	100.0%	100.0%	94.5%

Table 49: Percent of parameters recovered when Model 4 was the generating model and there were 30 observable variables

SS	LP probs	OP cond	% params rec for the LP levels				% params rec. for the OV			
			Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
100	1	1	99.0%	100.0%	100.0%	100.0%	95.9%	100.0%	98.8%	100.0%
100	1	2	97.0%	97.0%	98.0%	97.0%	95.7%	100.0%	99.1%	100.0%
100	2	1	92.0%	92.0%	92.0%	92.0%	96.7%	100.0%	95.4%	100.0%
100	2	2	92.0%	92.0%	92.0%	92.0%	96.9%	100.0%	95.3%	100.0%
500	1	1	93.0%	94.0%	93.0%	92.0%	95.9%	100.0%	98.0%	100.0%
500	1	2	97.0%	96.0%	96.0%	97.0%	95.8%	100.0%	97.9%	100.0%
500	2	1	95.0%	96.0%	95.0%	95.0%	96.2%	100.0%	97.2%	100.0%
500	2	2	92.0%	92.0%	91.0%	91.0%	95.6%	100.0%	90.7%	100.0%

When it came to fit, in the case where Model 1 was the generating model and there were 48 observables, with the small sample size the AIC and BIC tended to pick the constrained model that most closely fit with how the observable variable probability was structured (i.e. Model 2 was picked as the best fitting model when Model 1 was the generating model and the condition in which the probability of a correct response for the observables was based on a compensatory model was used. The DIC was split between this model and Model 1. When the sample size was large the BIC continued this pattern but the AIC and DIC shifted to pick Model 1 (see Table 50).

Table 50: Proportion of replications that each fit index picked each model for the best fitting model for the case where Model 1 was the generating model and there were 48 OVs

	LP	OP	AIC				BIC				DIC			
SS	probs	cond	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	0	1	0	0	0	1	0	0	0	1	0	0
100	1	2	0	0	1	0	0	0	1	0	0.7	0	0.3	0
100	1	3	0	0	0	1	0	0	0	1	0	0	0	1
100	2	1	0	1	0	0	0	1	0	0	0.9	0.1	0	0
100	2	2	0	0	1	0	0	0	1	0	1	0	0	0
100	2	3	0	0	0	1	0	0	0	1	0.2	0	0.1	0.7
500	1	1	1	0	0	0	0	1	0	0	1	0	0	0
500	1	2	1	0	0	0	0	0	1	0	1	0	0	0
500	1	3	1	0	0	0	0	0	0	1	1	0	0	0
500	2	1	1	0	0	0	0	1	0	0	1	0	0	0
500	2	2	1	0	0	0	0	0	1	0	1	0	0	0
500	2	3	1	0	0	0	0	0	0	1	1	0	0	0

When Model 1 was the generating model and 30 variables were used then the AIC and BIC tended to pick the constrained model in both the small and the large sample size cases. The DIC again shifted to pick Model 1 more often when the sample size was large (see Table 51).

Table 51: Proportion of replications that each fit index picked each model for the best fitting model for the case where Model 1 was the generating model and there were 30 OVs

	LP	OP	AIC				BIC				DIC			
SS	probs	cond	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	0	1	0	0	0	1	0	0	0.2	0.8	0	0
100	1	2	0	0	1	0	0	0	1	0	0.2	0	0.8	0
100	1	3	0	0	0	1	0	0	0	1	0.1	0	0.1	0.8
100	2	1	0	1	0	0	0	1	0	0	1	0	0	0
100	2	2	0	0	1	0	0	0	1	0	0.6	0	0.2	0.2
100	2	3	0	0	0	1	0	0	0	1	0.6	0	0.1	0.3
500	1	1	0	1	0	0	0	1	0	0	1	0	0	0
500	1	2	0	0	1	0	0	0	1	0	1	0	0	0
500	1	3	0	0	0	1	0	0	0	1	1	0	0	0
500	2	1	0	1	0	0	0	1	0	0	1	0	0	0
500	2	2	0	0	1	0	0	0	1	0	0.8	0	0	0.2
500	2	3	0	0	0	1	0	0	0	1	0.6	0	0.4	0

When Model 2 was the generating model then the AIC and the BIC picked Model 2 as the best fitting model. The DIC picked model 2 most of the time, although when the learning progressions were highly correlated it sometimes picked one of the other constrained model (see Table 52).

Table 52: Proportion of replications that each fit index picked each model for the best fitting model for the case where Model 2 was the generating model and there were 30 OVs

	LP	OP	AIC				BIC				DIC			
SS	probs	cond	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	0	1	0	0	0	1	0	0	0	1	0	0
100	1	2	0	1	0	0	0	1	0	0	0	1	0	0
100	2	1	0	1	0	0	0	1	0	0	0	0.5	0.3	0.2
100	2	2	0	1	0	0	0	1	0	0	0	0.9	0	0.1
500	1	1	0	1	0	0	0	1	0	0	0	1	0	0
500	1	2	0	1	0	0	0	1	0	0	0	1	0	0
500	2	1	0	1	0	0	0	1	0	0	0	0.9	0.1	0
500	2	2	0	1	0	0	0	1	0	0	0	0.5	0.4	0.1

The AIC and BIC picked Model 3 as the best fitting model when Model 3 was the generating model for both the small sample size cases and the large sample size cases.

The DIC shifted between Model 3 and Model 4 as the best fitting model in all cases (see Table 53).

Table 53: Proportion of replications that each fit index picked each model for the best fitting model for the case where Model 3 was the generating model and there were 30 OVs

	LP	OP	AIC				BIC				DIC			
SS	probs	cond	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	0	0	1	0	0	0	1	0	0	0	0.8	0.2
100	1	2	0	0	1	0	0	0	1	0	0	0	0.8	0.2
100	2	1	0	0	1	0	0	0	1	0	0	0	0.8	0.2
100	2	2	0	0	1	0	0	0	1	0	0	0	0.7	0.3
500	1	1	0	0	1	0	0	0	1	0	0	0	0.4	0.6
500	1	2	0	0	1	0	0	0	1	0	0	0	0.2	0.8
500	2	1	0	0	1	0	0	0	1	0	0	0	0.5	0.5
500	2	2	0	0	1	0	0	0	1	0	0	0	0.6	0.4

When Model 4 was the generating model then the AIC and the BIC picked Model 4 as the best fitting model. The DIC shifted between Model 4 and Model 3 as the best fitting model (see Table 54). In general, the constrained model that most closely matched how the data was generated was picked to be the best fitting model.

Table 54: Proportion of replications that each fit index picked each model for the best fitting model for the case where Model 4 was the generating model and there were 30 OVs

	LP	OP	AIC				BIC				DIC			
SS	probs	cond	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	0	0	0	1	0	0	0	1	0	0	0.1	0.9
100	1	2	0	0	0	1	0	0	0	1	0	0	0.1	0.9
100	2	1	0	0	0	1	0	0	0	1	0	0	0.1	0.9
100	2	2	0	0	0	1	0	0	0	1	0	0	0.2	0.8
500	1	1	0	0	0	1	0	0	0	1	0	0	0.8	0.2
500	1	2	0	0	0	1	0	0	0	1	0	0	0.8	0.2
500	2	1	0	0	0	1	0	0	0	1	0	0	0.6	0.4
500	2	2	0	0	0	1	0	0	0	1	0	0	0.3	0.7

When Model 1 was used to generate the data and 48 OVs were used, Model 1 had classification rates over 70% for each individual LP (see

Table 55 55) and over 69% for both LPs (see Table 56) when the same sample that was used to generate the data was used for classification. The minimum classification value occurred when the total attribute level required was dependent on the addition of the individual attribute levels (case 1). Also of note was that classification rates dropped from 20% to 30% when moving from the small sample size to the large sample size. This could be a similar “overfitting” effect as found in regression where sometimes small samples can take advantage of random chance assignment and appear to fit better than larger samples (Drasgow, Dorans, & Tucker, 1979).

Models 2, 3 and 4 had much lower classification rates, ranging from 25% to 66% for the individual LPs and from 7% to 47% for the combination of both LPs. Among these models each had the highest classification rate when the generated data was similar to the constraints within each model, although overall Model 2 seemed to outperform the other two models. In addition, the classification rates were higher when the two LPs were correlated.

The adjusted Rand statistic had a similar pattern (see Table 57 and Table 58), in that it indicated that Model 1 had a higher classification rate. However, the pattern between Models 2, 3 and 4 did not appear the same. For the classification of both LPs, Model 4 was the highest (from among these models) for most of the cells, but this pattern did not occur for each of the individual LPs.

Table 55: Classification rates for the individual LPs when Model 1 was used to generate the sample of 48 OVs and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on LP 1				correctly classified on LP 2			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	**86.6%**	39.9%	36.6%	35.5%	**85.7%**	39.4%	34.6%	35.1%
100	1	2	**94.4%**	42.7%	49.2%	35.4%	**94.6%**	42.1%	46.9%	34.7%
100	1	3	**86.1%**	45.5%	33.8%	48.4%	**87.3%**	46.3%	35.4%	49.3%
100	2	1	**91.2%**	52.2%	50.5%	42.0%	**91.0%**	53.4%	53.7%	51.9%
100	2	2	**95.5%**	51.8%	61.3%	53.1%	**94.3%**	47.3%	58.3%	35.2%
100	2	3	**92.5%**	56.7%	56.4%	66.5%	**91.6%**	54.0%	35.5%	61.8%
500	1	1	**73.1%**	35.5%	36.1%	31.4%	**72.6%**	36.2%	34.9%	34.4%
500	1	2	**84.1%**	41.5%	44.6%	27.1%	**84.3%**	38.5%	44.0%	29.8%
500	1	3	**84.7%**	45.1%	27.0%	47.7%	**85.6%**	43.9%	32.8%	46.0%
500	2	1	**84.2%**	49.4%	54.3%	45.2%	**83.7%**	48.4%	53.8%	49.5%
500	2	2	**89.6%**	44.8%	59.5%	28.6%	**89.4%**	45.9%	60.4%	35.0%
500	2	3	**91.7%**	47.5%	35.9%	65.4%	**92.7%**	49.9%	44.6%	63.7%

Table 56: Classification rates for the combination of LPs when Model 1 was used to generate the sample of 48 observable variables and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on both LPs			
			M1	M2	M3	M4
100	1	1	**85.2%**	20.2%	11.2%	11.0%
100	1	2	**91.5%**	16.7%	16.0%	10.3%
100	1	3	**80.6%**	18.6%	11.4%	14.0%
100	2	1	**89.0%**	35.5%	30.5%	22.0%
100	2	2	**92.5%**	28.8%	40.3%	18.9%
100	2	3	**85.9%**	35.8%	15.7%	45.7%
500	1	1	**69.4%**	12.9%	11.0%	7.4%
500	1	2	**74.1%**	11.7%	13.7%	7.3%
500	1	3	**73.1%**	14.5%	8.3%	12.7%
500	2	1	**80.0%**	35.2%	35.1%	24.7%
500	2	2	**83.9%**	26.7%	44.0%	8.3%
500	2	3	**86.6%**	31.8%	16.1%	46.9%

Table 57: The adjusted Rand statistic for the individual LPs when Model 1 was used to generate the sample of 48 observable variables and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	LP 1 classification				LP 2 classification			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	**0.687**	0.275	0.305	0.110	**0.498**	0.162	0.119	0.097
100	1	2	**0.864**	0.466	0.404	0.120	**0.530**	0.104	0.154	0.137
100	1	3	**0.675**	0.292	0.210	0.146	**0.573**	0.139	0.102	0.141
100	2	1	**0.791**	0.387	0.433	0.324	**0.716**	0.349	0.336	0.317
100	2	2	**0.893**	0.466	0.463	0.314	**0.678**	0.316	0.398	0.334
100	2	3	**0.816**	0.398	0.350	0.371	**0.705**	0.358	0.361	0.217
500	1	1	**0.445**	0.041	0.091	0.098	**0.519**	0.143	0.120	0.127
500	1	2	**0.638**	0.242	0.131	0.130	**0.521**	0.111	0.123	0.124
500	1	3	**0.637**	0.258	0.078	0.156	**0.565**	0.134	0.097	0.118
500	2	1	**0.655**	0.245	0.311	0.295	**0.705**	0.318	0.385	0.349
500	2	2	**0.772**	0.369	0.358	0.313	**0.720**	0.298	0.360	0.370
500	2	3	**0.808**	0.427	0.386	0.358	**0.754**	0.338	0.252	0.310

Table 58: The adjusted Rand statistic for the combination of LPs when Model 1 was used to generate the sample of 48 observable variables and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	both LPs classification			
			M1	M2	M3	M4
100	1	1	**0.561**	0.107	0.096	0.166
100	1	2	**0.523**	0.086	0.091	0.118
100	1	3	**0.557**	0.161	0.178	0.145
100	2	1	**0.791**	0.279	0.338	0.381
100	2	2	**0.777**	0.324	0.194	0.361
100	2	3	**0.798**	0.425	0.378	0.386
500	1	1	**0.565**	0.103	0.126	0.136
500	1	2	**0.512**	0.066	0.097	0.102
500	1	3	**0.514**	0.168	0.157	0.131
500	2	1	**0.780**	0.291	0.330	0.327
500	2	2	**0.774**	0.208	0.270	0.337
500	2	3	**0.809**	0.400	0.392	0.372

The classification rate for Model 1 decreases when the sample moves from the sample that was used to generate the parameters, to a separate sample that was generated in the same manner and with the same constraints as the previous sample (see Table 59 and Table 60). This decrease is particularly noticeable in the small sample case where the difference was as much as a 50% decrease. Model 1 still seemed have higher classification rates when the two LPs were correlated. In addition Model 2 did not consistently have a higher classification rate than the other constrained models. Instead Model 3 seemed to be highest in the cases where the data followed a compensatory or conjunctive model, while Model 4 was the best performing model (from among Models 2, 3, and 4) for the disjunctive model. While the Rand statistic did not always provide the exact same evidence regarding which model had better classifications (or better matching to the original classifications) it still showed the same pattern of drop off (see Table 61 and Table 62).

Table 59: Classification rates for the individual LPs when Model 1 was used to generate the sample of 48 observable variables and a separate sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on LP 1				correctly classified on LP 2			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	34.1%	32.8%	**34.8%**	34.3%	32.0%	**35.6%**	35.2%	31.0%
100	1	2	**63.2%**	37.2%	42.8%	29.1%	**61.7%**	39.0%	45.9%	25.8%
100	1	3	**60.2%**	39.3%	27.3%	44.4%	**64.2%**	39.6%	30.6%	44.6%
100	2	1	**67.0%**	48.4%	50.9%	37.2%	**69.2%**	48.9%	45.6%	46.2%
100	2	2	**75.2%**	46.0%	59.9%	45.6%	**73.6%**	42.1%	53.3%	29.2%
100	2	3	**83.5%**	51.0%	54.4%	60.4%	**82.9%**	47.7%	29.5%	59.4%
500	1	1	**34.2%**	33.4%	33.7%	30.2%	33.9%	34.3%	**34.8%**	31.6%
500	1	2	**73.1%**	40.1%	44.3%	25.7%	**71.4%**	35.6%	42.0%	28.4%
500	1	3	**75.6%**	42.9%	26.3%	46.0%	**76.5%**	42.7%	28.9%	46.3%
500	2	1	**67.6%**	50.0%	53.0%	45.5%	**67.7%**	47.8%	53.7%	47.4%
500	2	2	**81.5%**	42.1%	58.3%	27.9%	**82.5%**	43.9%	59.8%	33.9%
500	2	3	**86.3%**	46.3%	34.0%	64.3%	**87.3%**	48.3%	42.0%	63.4%

Table 60: Classification rates for the combination of LPs when Model 1 was used to generate the sample of 48 observable variables and a separate sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on both LPs			
			M1	M2	M3	M4
100	1	1	**23.6%**	14.6%	11.7%	9.9%
100	1	2	**44.4%**	13.6%	15.0%	6.7%
100	1	3	**42.6%**	13.4%	8.6%	11.1%
100	2	1	**61.7%**	31.2%	27.2%	17.1%
100	2	2	**64.0%**	25.4%	38.7%	13.7%
100	2	3	**72.3%**	32.3%	11.9%	39.8%
500	1	1	**27.3%**	11.4%	9.7%	6.4%
500	1	2	**56.2%**	10.8%	12.8%	6.2%
500	1	3	**59.0%**	13.5%	6.9%	12.7%
500	2	1	**60.8%**	35.1%	34.6%	24.4%
500	2	2	**72.1%**	26.0%	43.4%	7.5%
500	2	3	**77.4%**	30.9%	13.7%	46.1%

Table 61: The adjusted Rand statistic for the individual LPs when Model 1 was used to generate the sample of 48 observable variables and a separate sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	LP 1 classification				LP 2 classification			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	0.090	0.083	**0.164**	0.095	**0.407**	0.150	0.126	0.133
100	1	2	**0.312**	0.179	0.193	0.092	**0.435**	0.114	0.112	0.141
100	1	3	**0.281**	0.209	0.159	0.136	**0.419**	0.136	0.102	0.098
100	2	1	**0.404**	0.243	0.344	0.265	**0.596**	0.310	0.323	0.302
100	2	2	**0.543**	0.323	0.359	0.302	**0.577**	0.305	0.359	0.296
100	2	3	**0.645**	0.337	0.321	0.348	**0.593**	0.359	0.336	0.212
500	1	1	0.074	0.077	**0.133**	0.098	**0.416**	0.143	0.122	0.126
500	1	2	**0.441**	0.217	0.179	0.134	**0.408**	0.110	0.128	0.109
500	1	3	**0.469**	0.284	0.163	0.138	**0.460**	0.131	0.094	0.103
500	2	1	**0.401**	0.203	0.395	0.304	**0.616**	0.324	0.383	0.361
500	2	2	**0.629**	0.347	0.378	0.312	**0.619**	0.300	0.346	0.356
500	2	3	**0.695**	0.412	0.304	0.355	**0.652**	0.332	0.245	0.300

Table 62: The adjusted Rand statistic for the combination of LPs when Model 1 was used to generate the sample of 48 observable variables and a separate sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	both LPs classification			
			M1	M2	M3	M4
100	1	1	**0.465**	0.112	0.119	0.154
100	1	2	**0.421**	0.076	0.098	0.120
100	1	3	**0.449**	0.152	0.128	0.128
100	2	1	**0.678**	0.252	0.305	0.362
100	2	2	**0.638**	0.289	0.201	0.347
100	2	3	**0.698**	0.384	0.389	0.361
500	1	1	**0.462**	0.101	0.125	0.139
500	1	2	**0.409**	0.064	0.087	0.103
500	1	3	**0.415**	0.156	0.159	0.128
500	2	1	**0.679**	0.304	0.345	0.347
500	2	2	**0.675**	0.204	0.262	0.330
500	2	3	**0.694**	0.384	0.392	0.366

When Model 1 was used to generate the parameters and 30 observable variables were used the classification patterns were very similar as to the case when 48 observable variables was used. The classification rate was again lower in the case where there was a large sample size, the LPs were uncorrelated, and the OV probabilities were based on an additive model, at between 61% and 62%. The remaining classification rates for the individual LPs with Model 1 are between 70% and 94% (see Table 63). The classification rates for the combination of both LPs for Model 1 ranged from 76% to 89% for the small sample size and 51% to 72% for the large sample size (see

Table 64 64). The classification rates for Models 2-4 were at least 25% lower than that of Model 1, and were as low at 6% for the combination of both LPs. The pattern among Models 2-4, where they had the highest classification rates when the data was generated similar to their underlying model assumptions, continued.

The adjusted Rand statistic followed a similar pattern in that it was higher for Model 1 than for Models 2-4 for both the individual LPs (see Table 65) and the combination of LPs (see Table 66). This statistic also followed the pattern of indicating a better match for Model 1 when the sample size was small than when it was large.

Table 63: Classification rates for the individual LPs when Model 1 was used to generate the sample of 30 observable variables and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on LP 1				correctly classified on LP 2			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	**85.3%**	38.0%	36.1%	34.3%	**86.3%**	35.8%	35.8%	32.5%
100	1	2	**87.0%**	41.8%	44.0%	37.6%	**87.7%**	41.0%	43.2%	29.1%
100	1	3	**85.4%**	44.4%	32.6%	47.6%	**85.7%**	44.4%	34.0%	48.1%
100	2	1	**89.0%**	52.9%	47.1%	51.1%	**87.3%**	49.8%	44.9%	49.1%
100	2	2	**87.9%**	47.1%	53.1%	36.4%	**89.6%**	52.5%	57.0%	46.7%
100	2	3	**92.8%**	50.2%	45.1%	61.7%	**93.5%**	51.2%	43.9%	54.9%
500	1	1	**61.7%**	36.7%	34.9%	32.9%	**61.3%**	36.5%	34.3%	32.1%
500	1	2	**70.3%**	40.2%	43.8%	24.7%	**71.5%**	40.8%	43.9%	30.7%
500	1	3	**72.9%**	43.7%	31.6%	46.4%	**73.4%**	41.8%	25.1%	46.4%
500	2	1	**74.9%**	45.4%	48.3%	42.0%	**75.0%**	48.8%	47.6%	53.9%
500	2	2	**80.1%**	49.1%	59.9%	36.1%	**79.5%**	52.6%	58.7%	27.4%
500	2	3	**83.3%**	49.7%	29.2%	61.9%	**82.5%**	48.8%	36.7%	62.3%

Table 64: Classification rates for the combination of LPs when Model 1 was used to generate the sample of 30 observable variables and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on both LPs			
			M1	M2	M3	M4
100	1	1	**81.4%**	13.7%	11.3%	8.3%
100	1	2	**79.9%**	14.0%	12.9%	11.5%
100	1	3	**76.9%**	14.6%	10.2%	14.9%
100	2	1	**83.9%**	31.8%	21.2%	26.3%
100	2	2	**83.0%**	29.8%	34.1%	18.6%
100	2	3	**88.9%**	30.9%	20.0%	37.4%
500	1	1	**52.7%**	12.6%	10.1%	8.5%
500	1	2	**51.2%**	11.8%	12.7%	7.8%
500	1	3	**52.6%**	12.3%	6.8%	14.1%
500	2	1	**66.6%**	33.9%	26.1%	26.9%
500	2	2	**68.4%**	35.2%	42.9%	10.2%
500	2	3	**71.7%**	32.1%	11.4%	46.1%

Table 65: The adjusted Rand statistic for the individual LPs when Model 1 was used to generate the sample of 30 observable variables and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	LP 1 classification				LP 2 classification			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	**0.665**	0.280	0.226	0.099	**0.472**	0.097	0.107	0.088
100	1	2	**0.693**	0.306	0.206	0.123	**0.498**	0.078	0.119	0.107
100	1	3	**0.652**	0.261	0.129	0.144	**0.548**	0.095	0.057	0.095
100	2	1	**0.749**	0.312	0.314	0.308	**0.668**	0.263	0.263	0.219
100	2	2	**0.724**	0.358	0.311	0.252	**0.708**	0.290	0.279	0.313
100	2	3	**0.830**	0.447	0.414	0.284	**0.697**	0.294	0.194	0.266
500	1	1	0.282	0.275	**0.286**	0.105	**0.403**	0.102	0.106	0.105
500	1	2	**0.390**	0.312	0.164	0.117	**0.418**	0.082	0.115	0.118
500	1	3	**0.432**	0.337	0.176	0.143	**0.434**	0.094	0.099	0.046
500	2	1	**0.495**	0.398	0.431	0.296	**0.603**	0.273	0.284	0.304
500	2	2	**0.586**	0.474	0.450	0.305	**0.611**	0.271	0.325	0.321
500	2	3	**0.633**	0.516	0.476	0.323	**0.620**	0.287	0.172	0.228

Table 66: The adjusted Rand statistic for the individual LPs when Model 1 was used to generate the sample of 30 observable variables and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	both LPs classification			
			M1	M2	M3	M4
100	1	1	**0.407**	0.098	0.082	0.102
100	1	2	**0.384**	0.110	0.043	0.079
100	1	3	**0.380**	0.152	0.158	0.102
100	2	1	**0.552**	0.272	0.269	0.271
100	2	2	**0.543**	0.205	0.257	0.268
100	2	3	**0.599**	0.333	0.306	0.268
500	1	1	**0.404**	0.107	0.096	0.092
500	1	2	**0.388**	0.041	0.077	0.071
500	1	3	**0.363**	0.144	0.137	0.095
500	2	1	**0.572**	0.266	0.313	0.296
500	2	2	**0.571**	0.232	0.161	0.267
500	2	3	**0.524**	0.340	0.341	0.288

When the sample used for classification rates was a separate sample than the one used for generating the parameters then the classification rate dropped for Model 1. However, the classification rates were similar between the two samples for Models 2, 3 and 4 (see Table 67 and Table 68). The inflation of the classification rates for the small sample size did not appear with this sample. The adjusted Rand statistics were low for this sample (see Table 69 and Table 70). The highest value for the match between the combination of the LPs was .336 and the highest value was .484 for the individual LPs.

Table 67: Classification rates for the individual LPs when Model 1 was used to generate the sample of 30 observable variables and a separate sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on LP 1				correctly classified on LP 2			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	32.8%	33.6%	32.7%	**34.3%**	30.6%	30.9%	**32.3%**	28.9%
100	1	2	**49.6%**	35.8%	40.4%	32.9%	**47.6%**	35.7%	40.8%	24.4%
100	1	3	**50.5%**	36.5%	26.4%	43.7%	**51.7%**	38.6%	27.4%	43.6%
100	2	1	**61.1%**	47.6%	43.5%	43.9%	**61.4%**	47.3%	41.4%	48.5%
100	2	2	**64.7%**	41.0%	50.4%	26.0%	**65.8%**	47.7%	52.8%	42.1%
100	2	3	**71.6%**	41.0%	39.2%	56.1%	**74.3%**	45.2%	37.8%	53.9%
500	1	1	33.8%	**35.7%**	34.4%	32.8%	33.7%	**35.2%**	33.6%	31.9%
500	1	2	**57.9%**	38.9%	42.9%	23.1%	**57.7%**	40.3%	43.5%	28.3%
500	1	3	**61.3%**	42.8%	30.2%	48.2%	**60.8%**	40.7%	23.9%	45.2%
500	2	1	**58.3%**	45.2%	47.7%	40.1%	**59.0%**	48.9%	46.5%	55.2%
500	2	2	**71.2%**	49.0%	58.2%	33.9%	**70.8%**	50.6%	57.7%	26.6%
500	2	3	**75.0%**	46.7%	27.3%	59.6%	**74.6%**	47.1%	36.4%	61.2%

Table 68: Classification rates for the combination of LPs when Model 1 was used to generate the sample of 30 observable variables and a separate sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on both LPs			
			M1	M2	M3	M4
100	1	1	**18.4%**	11.3%	9.6%	9.8%
100	1	2	**22.9%**	9.4%	10.6%	7.0%
100	1	3	**23.6%**	11.3%	6.6%	12.1%
100	2	1	**51.3%**	29.3%	16.1%	24.7%
100	2	2	**48.6%**	26.5%	30.5%	11.7%
100	2	3	**57.6%**	25.3%	13.7%	34.6%
500	1	1	**22.2%**	12.1%	10.3%	9.2%
500	1	2	**30.2%**	11.8%	13.3%	6.2%
500	1	3	**32.8%**	12.1%	6.2%	14.3%
500	2	1	**47.8%**	33.6%	25.3%	26.2%
500	2	2	**54.4%**	35.0%	42.3%	9.0%
500	2	3	**58.4%**	30.0%	10.0%	43.7%

Table 69: The adjusted Rand statistic for the individual LPs when Model 1 was used to generate the sample of 30 observable variables and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	LP 1 classification				LP 2 classification			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	0.088	0.056	**0.160**	0.106	0.073	0.103	**0.106**	0.082
100	1	2	**0.170**	0.158	0.143	0.104	0.086	0.079	0.102	**0.097**
100	1	3	0.171	**0.202**	0.154	0.107	**0.121**	0.088	0.055	0.065
100	2	1	0.338	0.345	**0.401**	0.293	0.258	**0.271**	0.265	0.226
100	2	2	0.360	**0.381**	**0.381**	0.238	0.269	0.262	0.270	**0.279**
100	2	3	0.433	**0.471**	0.470	0.274	**0.302**	0.282	0.166	0.263
500	1	1	0.076	0.078	**0.158**	0.108	0.108	0.109	0.107	**0.118**
500	1	2	**0.235**	0.230	0.172	0.118	**0.119**	0.083	0.112	0.114
500	1	3	**0.281**	0.274	0.192	0.151	**0.128**	0.099	0.101	0.046
500	2	1	0.284	0.295	**0.413**	0.299	**0.315**	0.282	0.281	0.299
500	2	2	0.440	0.432	**0.459**	0.308	0.304	0.274	**0.312**	0.311
500	2	3	**0.484**	0.479	0.440	0.302	**0.314**	0.280	0.168	0.225

Table 70: The adjusted Rand statistic for the individual LPs when Model 1 was used to generate the sample of 30 observable variables and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	both LPs classification			
			M1	M2	M3	M4
100	1	1	0.102	**0.106**	0.080	0.096
100	1	2	0.079	**0.099**	0.037	0.083
100	1	3	0.068	0.119	**0.136**	0.103
100	2	1	0.270	0.255	0.263	**0.297**
100	2	2	0.226	0.213	0.231	**0.263**
100	2	3	0.297	0.310	**0.311**	0.267
500	1	1	0.109	0.110	**0.117**	0.098
500	1	2	**0.084**	0.041	0.074	0.068
500	1	3	0.069	**0.154**	0.129	0.104
500	2	1	0.272	0.257	**0.318**	0.299
500	2	2	**0.272**	0.237	0.158	0.269
500	2	3	0.237	0.316	**0.336**	0.276

When the generated data was based on Model 2 and the sample used for classification was the same as the sample used for generating the data, the same issue, as when the data was based on Model 1, of having higher classification rates for Model 1 with the small sample size than with the large sample size, occurred (see Table 71). In addition, except for two cases with Model 1 and the low sample size, the classification rates were lower for the cases where the LPs were uncorrelated than when the LPs were correlated. When the LPs were uncorrelated the classification rate for Models 2-4 was in the high 30% low 40% range, and when the LPs were correlated this range was in the high 40% to low 50% range (see Table 71). For the large sample size Model 1 had classification rates about 10% higher than the other models. Among Models 2-4, Model 2 had the highest classification rate.

The classification rates for this sample with regards to the combination of LPs were lower with a range of 10% to 15% for Models 2-4 when the LPs were uncorrelated and 20%-30% when the LPs were correlated. The classification rate for Model 1 under the large sample size condition was in the low 30% when the LPs were uncorrelated and the high 40% when the LPs were correlated (see Table 72). Model 2 had the highest classification rate among Models 2-4 when the LPs were correlated. The classification rates were all very similar when the LPs were uncorrelated.

Table 71: Classification rates for the individual LPs when Model 2 was used to generate the data and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on LP 1				correctly classified on LP 2			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	**79.2%**	41.7%	39.8%	38.4%	**77.1%**	41.5%	37.5%	34.5%
100	1	2	**77.9%**	39.4%	34.5%	36.3%	**79.7%**	42.3%	39.2%	39.1%
100	2	1	**75.8%**	52.2%	45.6%	50.2%	**76.2%**	52.3%	46.2%	41.2%
100	2	2	**81.0%**	53.7%	47.6%	46.8%	**81.2%**	54.3%	49.4%	48.6%
500	1	1	**53.5%**	40.8%	36.7%	37.0%	**53.7%**	41.1%	36.4%	35.4%
500	1	2	**53.4%**	41.0%	35.2%	36.8%	**53.0%**	41.9%	37.5%	36.4%
500	2	1	**61.6%**	52.3%	52.1%	50.7%	**61.7%**	51.3%	47.0%	47.1%
500	2	2	**62.7%**	53.2%	48.8%	46.0%	**62.3%**	53.3%	52.9%	50.7%

Table 72: Classification rates for the combination of LPs when Model 2 was used to generate the data and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on both LPs			
			M1	M2	M3	M4
100	1	1	**67.9%**	14.2%	14.4%	12.7%
100	1	2	**69.0%**	11.6%	10.9%	12.5%
100	2	1	**68.2%**	36.5%	23.1%	24.3%
100	2	2	**73.5%**	36.8%	25.3%	24.6%
500	1	1	**31.2%**	11.7%	12.1%	11.9%
500	1	2	**30.5%**	11.9%	11.9%	12.5%
500	2	1	**47.3%**	36.6%	30.1%	30.2%
500	2	2	**48.9%**	38.1%	33.8%	32.1%

The adjusted Rand statistic showed similar patterns to the classification rates; the individual LPs had a higher rate in some cases for the low sample size than for the high sample size (see Table 73), although this was not seen in the statistic for the combination of both LPs (see Table 74). The pattern where there seemed to be a better match when the LPs were correlated than when they were uncorrelated was also reflected in the adjusted Rand statistic.

Table 73: The adjusted Rand statistic for the individual LPs when Model 2 was used to generate the data and the same sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	LP 1 classification				LP 2 classification			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	**0.546**	0.494	0.395	0.123	**0.116**	0.090	0.112	0.104
100	1	2	0.509	**0.551**	0.419	0.098	**0.111**	0.092	0.087	0.106
100	2	1	**0.515**	0.512	0.487	0.281	**0.275**	0.257	0.238	0.212
100	2	2	**0.599**	**0.599**	0.569	0.296	**0.292**	0.269	0.209	0.271
500	1	1	0.210	**0.212**	0.130	0.133	**0.138**	0.094	0.112	0.109
500	1	2	**0.208**	0.203	0.133	0.139	**0.135**	0.100	0.104	0.115
500	2	1	0.359	0.345	**0.387**	0.295	**0.286**	0.254	0.282	0.246
500	2	2	0.370	0.371	**0.427**	0.309	**0.314**	0.277	0.258	0.310

Table 74: The adjusted Rand statistic for the combination of LPs when Model 2 was used to generate the data and the same sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	both LPs classification			
			M1	M2	M3	M4
100	1	1	0.079	**0.107**	0.073	0.074
100	1	2	0.094	0.080	**0.098**	0.083
100	2	1	0.235	**0.243**	0.184	0.246
100	2	2	**0.249**	0.197	0.248	0.226
500	1	1	0.084	**0.116**	0.104	0.084
500	1	2	0.083	**0.113**	0.110	0.084
500	2	1	0.255	**0.274**	0.224	0.248
500	2	2	**0.290**	0.244	0.293	0.286

When the sample that was being classified was not the same as the sample used to generate the models, then the classification rates for the small sample size was very similar to that of the large sample size for all models. The classification rates for all models were very similar to each other. When the LPs were not correlated the classification rate for the individual LP had values between 29% and 41%, while the values when the LPs were correlated were between 37% and 53% (see Table 75). The classification rates dropped for the combination of both LPs with the rate being in-between 8% and 14% when the LPs were uncorrelated and between 21% and 38% when the LPs were correlated (see Table 76).

The adjusted Rand statistic did not indicate a good match for any of the models, although it also displayed the pattern of better matching when the LPs were correlated than when they were uncorrelated. The values were slightly higher for the individual LPs (see Table 77) than for the combination of LPs (see Table 78) but in either case the highest value was below .4 indicating that there was not a good match between the classifications.

Table 75: Classification rates for the individual LPs when Model 2 was used to generate the sample of 30 observable variables and a separate sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on LP 1				correctly classified on LP 2			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	34.6%	**39.5%**	37.1%	32.8%	32.9%	**36.2%**	33.1%	31.3%
100	1	2	30.1%	**38.2%**	29.7%	34.0%	33.9%	**40.8%**	35.5%	36.2%
100	2	1	49.2%	**52.5%**	44.6%	47.8%	49.3%	**50.9%**	43.1%	37.2%
100	2	2	50.4%	**53.3%**	43.4%	43.1%	51.8%	**53.1%**	45.0%	46.3%
500	1	1	35.6%	**39.2%**	35.7%	35.7%	37.3%	**41.2%**	35.5%	34.7%
500	1	2	37.0%	**40.8%**	34.9%	35.6%	36.9%	**40.1%**	36.0%	35.5%
500	2	1	51.5%	**52.4%**	50.3%	49.1%	49.5%	**52.6%**	45.5%	47.4%
500	2	2	48.5%	**51.5%**	47.7%	44.8%	49.8%	**50.8%**	51.2%	49.0%

Table 76: Classification rates for the combination of LPs when Model 2 was used to generate the sample of 30 observable variables and a separate sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on both LPs			
			M1	M2	M3	M4
100	1	1	**12.1%**	11.4%	11.8%	9.1%
100	1	2	9.3%	**12.4%**	8.8%	10.2%
100	2	1	37.0%	**38.0%**	23.8%	22.7%
100	2	2	36.4%	**37.6%**	21.9%	23.6%
500	1	1	**13.0%**	11.6%	11.7%	11.3%
500	1	2	**13.3%**	12.3%	12.0%	12.6%
500	2	1	35.1%	**37.7%**	29.0%	29.9%
500	2	2	33.3%	**37.1%**	33.3%	31.5%

Table 77: The adjusted Rand statistic for the individual LPs when Model 2 was used to generate the sample of 30 observable variables and a separate sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	LP 1 classification				LP 2 classification			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	0.072	**0.076**	0.074	0.132	**0.119**	0.096	0.104	0.099
100	1	2	0.043	**0.094**	0.079	0.116	**0.124**	0.094	0.078	0.118
100	2	1	0.269	0.292	**0.352**	0.306	0.303	**0.305**	0.243	0.222
100	2	2	0.276	0.296	**0.365**	0.314	**0.331**	0.294	0.197	0.276
500	1	1	0.090	0.110	0.103	**0.123**	**0.146**	0.095	0.100	0.114
500	1	2	0.106	0.100	0.103	**0.140**	**0.124**	0.095	0.101	0.103
500	2	1	0.289	0.266	**0.389**	0.311	**0.319**	0.285	0.285	0.264
500	2	2	0.261	0.260	**0.377**	0.313	**0.300**	0.279	0.266	0.296

Table 78: The adjusted Rand statistic for the combination of both LPs when Model 2 was used to generate the sample, and a separate sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	both LPs classification			
			M1	M2	M3	M4
100	1	1	**0.098**	0.095	0.080	0.081
100	1	2	0.091	0.102	**0.106**	0.090
100	2	1	0.269	0.269	0.169	**0.276**
100	2	2	**0.274**	0.178	0.263	0.241
500	1	1	0.080	0.105	**0.109**	0.090
500	1	2	0.083	**0.109**	0.095	0.081
500	2	1	0.278	**0.284**	0.243	0.275
500	2	2	**0.291**	0.248	0.282	0.284

When data was generated using Model 3 and the same sample was used for classification as for parameter generation Model 1 again had higher classification rates than the other models and the classification rates for the small sample size for Model 1 was higher than for the large sample size. Model 4 had the lowest classification rate for the individual LPs (see Table 79) and for most of the cells with the combination of LPs (see Table 80). In addition the classification rates were again higher for the cases when the LPs were correlated than when they were not correlated.

Table 79: Classification rates for the individual LPs when Model 3 was used to generate the sample, and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on LP 1				correctly classified on LP 2			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	**75.2%**	33.1%	34.9%	27.5%	**72.3%**	36.7%	35.6%	33.8%
100	1	2	**79.2%**	38.0%	37.9%	30.6%	**79.1%**	39.4%	39.2%	33.8%
100	2	1	**69.6%**	42.5%	42.4%	31.0%	**71.9%**	45.9%	44.4%	38.1%
100	2	2	**75.6%**	45.1%	43.9%	35.0%	**76.4%**	46.9%	46.8%	37.6%
500	1	1	**47.7%**	35.3%	35.5%	24.1%	**47.3%**	35.0%	35.7%	28.4%
500	1	2	**47.9%**	34.9%	35.9%	26.7%	**47.8%**	36.1%	36.5%	24.2%
500	2	1	**53.4%**	44.5%	43.9%	26.0%	**52.7%**	44.4%	44.0%	33.9%
500	2	2	**54.3%**	45.2%	45.4%	28.6%	**54.1%**	44.6%	44.7%	34.5%

Table 80: Classification rates for the combination of LPs when Model 3 was used to generate the sample, and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	correctly classified on both LPs			
			M1	M2	M3	M4
100	1	1	**62.8%**	10.3%	10.4%	10.1%
100	1	2	**70.0%**	12.9%	11.5%	11.9%
100	2	1	**62.8%**	29.0%	24.2%	13.0%
100	2	2	**69.8%**	27.9%	26.9%	13.1%
500	1	1	**24.8%**	9.5%	9.9%	6.7%
500	1	2	**25.8%**	10.0%	10.0%	6.6%
500	2	1	**36.7%**	33.6%	26.5%	10.0%
500	2	2	**38.2%**	30.4%	28.3%	7.2%

The adjusted Rand statistic when the data was generated based on Model 3 and the same parameters were used for generating the parameters and classification followed a similar pattern as the classification rates. Model 1 had higher adjusted Rand statistic in the small sample size than the large sample size (see Table 81). The adjusted Rand statistic also indicated a better match when the LPs were correlated than when they were not correlated for both the individual LPs and for the combination of LPs. The adjusted Rand statistic was very low for the combination of LPs with a maximum value of .161 (see Table 82).

Table 81: The adjusted Rand statistic for the individual LPs when Model 3 was used to generate the sample, and the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	LP 1 classification				LP 2 classification			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	**0.454**	0.410	0.301	0.065	**0.094**	0.047	0.066	0.090
100	1	2	0.531	0.538	**0.415**	0.096	**0.115**	0.061	0.092	0.111
100	2	1	0.388	**0.415**	0.367	0.156	**0.195**	0.179	0.187	0.192
100	2	2	0.498	**0.505**	0.478	0.179	**0.208**	0.180	0.177	0.205
500	1	1	**0.139**	0.133	0.076	0.089	**0.088**	0.052	**0.088**	**0.088**
500	1	2	**0.144**	0.142	0.084	0.090	0.094	0.057	0.091	**0.095**
500	2	1	0.218	**0.221**	0.211	0.183	0.183	0.162	0.178	**0.191**
500	2	2	**0.237**	0.224	0.229	0.201	0.189	0.185	**0.197**	0.188

Table 82: The adjusted Rand statistic for the combination of LPs when Model 3 was used to generate the sample, the same sample was used to generate the parameters as to estimate classification.

SS	LP probs	OP cond	both LPs classification			
			M1	M2	M3	M4
100	1	1	0.041	0.031	**0.088**	0.037
100	1	2	0.059	0.052	**0.086**	0.053
100	2	1	0.150	0.086	**0.151**	0.144
100	2	2	0.151	0.107	**0.161**	0.136
500	1	1	0.050	0.019	**0.057**	0.033
500	1	2	**0.053**	0.050	0.032	0.034
500	2	1	**0.155**	0.076	0.119	0.139
500	2	2	**0.159**	0.099	0.129	0.157

When a separate sample was used, the classification rates for the individual LPs were fairly similar to those of the previous sample for Models 2-4, but were lower for Model 1 (see Table 83). The highest classification rate was in either Model 2 or Model 3, although it was never higher than 45%. For the combination of LPs the cases in which the LPs were correlated had higher classification rates than the cases in which the LPS were not correlated, but was not higher than 33.1% in any case (see Table 84).

Table 83: Classification rates for the individual LPs when Model 3 was used to generate the sample, and a separate sample was used to generate the parameters as to estimate classification rates

SS	LP probs	OP cond	correctly classified on LP 1				correctly classified on LP 2			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	24.2%	31.8%	**32.1%**	22.5%	28.6%	31.4%	**31.6%**	28.1%
100	1	2	29.4%	33.4%	**34.5%**	23.4%	28.5%	32.9%	**34.2%**	25.8%
100	2	1	38.9%	40.9%	**42.1%**	26.1%	37.0%	**43.9%**	38.7%	37.3%
100	2	2	39.9%	40.4%	**43.0%**	30.0%	37.6%	40.8%	**41.6%**	31.1%
500	1	1	32.0%	**34.6%**	34.3%	21.9%	32.4%	33.5%	**34.1%**	27.1%
500	1	2	31.8%	34.4%	**34.7%**	24.5%	32.0%	**33.8%**	33.6%	23.1%
500	2	1	41.7%	**43.9%**	42.8%	24.7%	41.6%	**43.5%**	42.7%	32.5%
500	2	2	39.9%	**45.1%**	44.5%	28.3%	40.9%	**43.5%**	**43.5%**	33.8%

Table 84: Classification rates for the combination of LPs when Model 3 was used to generate the sample, and a separate sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	correctly classified on both LPs			
			M1	M2	M3	M4
100	1	1	5.8%	7.1%	6.5%	**7.4%**
100	1	2	8.5%	**9.9%**	9.6%	7.5%
100	2	1	23.4%	**28.8%**	22.1%	12.7%
100	2	2	24.4%	23.6%	**25.3%**	9.3%
500	1	1	8.8%	**9.4%**	9.3%	5.8%
500	1	2	9.2%	**9.4%**	9.1%	6.2%
500	2	1	24.3%	**33.1%**	26.6%	9.2%
500	2	2	22.4%	**30.2%**	27.9%	6.8%

The adjusted Rand statistic also indicated a poor match between the original classification and the classification from the models, as it was below .2 for the individual LPs (see Table 85) as well as the combination of LPs (see Table 86). For both the individual LPs and the combination of LPs the pattern of the Rand statistic being lower for the cases in which the LPs were uncorrelated was found here as well.

Table 85: The adjusted Rand statistic for the individual LPs when Model 3 was used to generate the sample, and a separate sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	LP 1 classification				LP 2 classification			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	0.021	0.040	0.028	**0.073**	0.075	0.058	**0.079**	0.064
100	1	2	0.037	0.036	0.038	**0.099**	0.090	0.064	**0.099**	0.093
100	2	1	0.112	0.111	0.120	**0.192**	**0.210**	0.196	0.205	0.207
100	2	2	0.127	0.103	0.149	**0.162**	0.162	0.156	0.165	**0.172**
500	1	1	0.061	0.066	0.056	**0.097**	0.088	0.055	**0.089**	**0.089**
500	1	2	0.065	0.064	0.056	**0.091**	0.090	0.057	**0.093**	0.088
500	2	1	0.146	0.151	0.180	**0.193**	0.180	0.166	0.182	**0.183**
500	2	2	0.135	0.134	0.170	**0.194**	0.182	0.172	**0.198**	0.184

Table 86: The adjusted Rand statistic for the combination of LPs when Model 3 was used to generate the sample, and a separate sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	both LPs classification			
			M1	M2	M3	M4
100	1	1	0.044	0.032	**0.065**	0.047
100	1	2	0.056	0.044	**0.060**	0.052
100	2	1	0.167	0.074	**0.194**	0.180
100	2	2	**0.147**	0.084	0.128	0.127
500	1	1	0.051	0.017	**0.054**	0.036
500	1	2	**0.052**	0.048	0.028	0.033
500	2	1	**0.155**	0.072	0.112	0.139
500	2	2	**0.160**	0.089	0.138	0.159

Much of the same patterns can be seen when the data was generated following Model 4. When the same sample was used for generating the data as for classifying students then Model 1 had a high classification rate in the small sample size for both the individual LPs (see Table 87) as well as for the combination of LPs (see Table 88). For the individual LPs as well as the combination of LPs, Model 3 had the lowest classification rate.

Table 87: Classification rates for the individual LPs when Model 4 was used to generate the sample, and the same sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	correctly classified on LP 1				correctly classified on LP 2			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	**76.7%**	37.5%	31.2%	37.0%	**77.7%**	37.3%	29.8%	38.1%
100	1	2	**77.8%**	34.2%	28.5%	35.8%	**76.9%**	37.7%	31.9%	35.6%
100	2	1	**74.5%**	45.6%	39.2%	45.5%	**74.5%**	43.9%	34.3%	42.9%
100	2	2	**72.7%**	46.9%	37.3%	44.5%	**73.5%**	48.2%	38.9%	46.3%
500	1	1	**47.3%**	34.7%	25.3%	35.1%	**47.9%**	36.3%	24.3%	36.9%
500	1	2	**49.1%**	35.7%	26.7%	36.9%	**48.8%**	36.2%	23.8%	37.3%
500	2	1	**53.0%**	44.5%	32.0%	44.8%	**53.9%**	44.0%	25.2%	45.0%
500	2	2	**53.4%**	44.4%	33.9%	45.2%	**54.6%**	45.0%	33.6%	45.3%

Table 88: Classification rates for the combination of LPs when Model 4 was used to generate the sample, and the same sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	correctly classified on both LPs			
			M1	M2	M3	M4
100	1	1	**67.0%**	10.7%	9.4%	10.9%
100	1	2	**68.5%**	8.4%	8.5%	9.1%
100	2	1	**67.1%**	27.3%	13.3%	25.9%
100	2	2	**63.9%**	31.5%	15.2%	26.3%
500	1	1	**25.6%**	9.9%	6.7%	10.4%
500	1	2	**26.4%**	10.4%	6.6%	10.4%
500	2	1	**37.8%**	32.2%	7.5%	29.7%
500	2	2	**37.6%**	32.5%	13.0%	28.3%

The adjusted Rand statistic indicated a poor match between the original classifications and the classifications indicated from the application of each model. While there were some cases for LP1 when the sample size was small that had values around .5 (see Table 89), for LP2 and the high sample size case the highest value was .23. For the combination of LPs all adjusted Rand statistics were below .2 (see Table 90).

Table 89: The adjusted Rand statistic for the individual LPs when Model 4 was used to generate the sample, and the same sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	LP 1 classification				LP 2 classification			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	0.490	**0.510**	0.374	0.089	**0.087**	0.052	0.063	0.040
100	1	2	**0.507**	0.491	0.389	0.081	**0.103**	0.055	0.051	0.067
100	2	1	**0.471**	**0.471**	0.433	0.183	0.160	**0.177**	0.124	0.106
100	2	2	0.445	**0.455**	0.413	0.208	**0.207**	0.200	0.129	0.128
500	1	1	0.129	**0.151**	0.080	0.077	**0.102**	0.053	0.030	0.040
500	1	2	**0.152**	0.150	0.087	0.096	**0.097**	0.061	0.057	0.033
500	2	1	**0.222**	0.219	0.213	0.187	**0.182**	0.165	0.110	0.059
500	2	2	0.219	**0.237**	0.230	0.175	**0.197**	0.178	0.098	0.114

Table 90: The adjusted Rand statistic for the combination of LPs when Model 4 was used to generate the sample, and the same sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	both LPs classification			
			M1	M2	M3	M4
100	1	1	0.045	0.088	**0.091**	0.050
100	1	2	0.047	0.083	**0.086**	0.051
100	2	1	0.155	**0.170**	0.169	0.157
100	2	2	0.160	0.175	**0.205**	0.155
500	1	1	0.030	0.077	**0.101**	0.052
500	1	2	0.039	0.096	**0.098**	0.058
500	2	1	0.128	**0.187**	0.179	0.152
500	2	2	0.162	0.181	**0.192**	0.151

When a separate sample was used to classify the students then the pattern of lower classification rates when the LPs were correlated was found both in with the classifications for the individual LPs and the combination of LPs. The classification rates for the individual LPs were all under 45% and were lowest for Model 3 (see Table 91). The highest classification rates for the combination of LPs were found in the correlated LP case under Models 1, 2, and 4 with the rates varying from 22% to 33% (see Table 92). The highest rate among the other cases was 11.1% for the combination of LPs.

Table 91: Classification rates when Model 4 was used to generate the sample of 30 observable variables and a separate sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	correctly classified on LP 1				correctly classified on LP 2			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	28.1%	33.4%	30.9%	**33.5%**	27.4%	**33.8%**	24.1%	33.2%
100	1	2	25.5%	**32.4%**	22.4%	32.1%	28.2%	**35.4%**	28.2%	34.1%
100	2	1	36.3%	35.9%	31.2%	**39.3%**	39.1%	**43.1%**	29.6%	41.9%
100	2	2	36.7%	**42.3%**	33.2%	38.7%	40.4%	**44.5%**	30.0%	40.5%
500	1	1	31.0%	**35.2%**	22.9%	**35.2%**	33.0%	34.7%	23.9%	**35.5%**
500	1	2	32.4%	34.1%	25.2%	**34.5%**	31.2%	34.4%	22.5%	**34.8%**
500	2	1	41.2%	44.0%	31.8%	**44.4%**	40.3%	**44.4%**	23.3%	**44.4%**
500	2	2	40.4%	**44.3%**	31.7%	44.0%	40.3%	**44.8%**	32.2%	43.8%

Table 92: Classification rates when Model 4 was used to generate the sample of 30 observable variables and a separate sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	correctly classified on both LPs			
			M1	M2	M3	M4
100	1	1	8.0%	8.2%	**9.3%**	8.2%
100	1	2	7.8%	**8.2%**	6.4%	6.7%
100	2	1	**24.2%**	22.9%	8.0%	22.0%
100	2	2	23.1%	**27.9%**	9.5%	21.2%
500	1	1	9.4%	**10.1%**	5.7%	10.0%
500	1	2	8.6%	9.3%	6.0%	**9.6%**
500	2	1	22.9%	**32.8%**	7.3%	30.2%
500	2	2	22.4%	**32.6%**	11.1%	27.8%

The adjusted Rand statistic again indicated a poor match with again all values under .2 for both the individual LPs (see Table 93) and the combination of LPs (see

Table 94 94). It also indicated a better match when the LPs were correlated.

Table 93: The adjusted Rand statistic for the individual LPs when Model 4 was used to generate the sample, and a separate sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	LP 1 classification				LP 2 classification			
			M1	M2	M3	M4	M1	M2	M3	M4
100	1	1	0.040	0.030	0.036	**0.087**	**0.080**	0.053	0.063	0.034
100	1	2	0.022	0.039	0.034	**0.076**	**0.118**	0.067	0.033	0.069
100	2	1	0.094	0.108	0.111	**0.158**	**0.180**	0.161	0.101	0.090
100	2	2	0.100	0.130	0.130	**0.174**	**0.204**	0.174	0.094	0.108
500	1	1	0.058	0.072	0.055	**0.095**	**0.094**	0.056	0.030	0.044
500	1	2	0.068	0.061	0.062	**0.090**	**0.091**	0.061	0.051	0.029
500	2	1	0.150	0.129	0.171	**0.197**	**0.186**	0.176	0.113	0.058
500	2	2	0.137	0.133	0.166	**0.191**	**0.190**	0.181	0.096	0.108

Table 94: The adjusted Rand statistic for the combination of LPs when Model 4 was used to generate the sample, and a separate sample was used to generate the parameters as to estimate classification

SS	LP probs	OP cond	both LPs classification			
			M1	M2	M3	M4
100	1	1	0.044	**0.084**	0.077	0.048
100	1	2	0.051	0.082	**0.107**	0.065
100	2	1	0.141	0.147	**0.187**	0.138
100	2	2	0.162	0.170	**0.199**	0.165
500	1	1	0.031	0.092	**0.095**	0.055
500	1	2	0.038	**0.092**	0.089	0.057
500	2	1	0.133	**0.192**	0.185	0.161
500	2	2	0.159	**0.186**	0.175	0.150

Discussion

While the constrained model most closely associated with the generating model was most often picked to be the best fitting model, Model 2 and Model 1 seemed to be able to best reproduce the generating LP and OV probabilities overall. All models were able to reproduce their own generating parameters, but Models 3 and Model 4 had lower rates of parameter recovery when a different model was used for data generation. This makes sense, particularly for the case where Model 3 was used for generating the data and Model 4 was used to generate the parameters (and vice versa). For Models 3 and 4 the probability of a correct response only depends on students' ability on one of the LPs. Model 3 depends on the student's ability for the LP in which they have the lower ability, while Model 4 depends on the student's ability for the LP in which they have the higher ability, which means that these models depend on the ability level for the opposite LP as each other, therefore it would make sense that they would have difficulty estimating each other's probabilities, or probabilities that would depend on the ability level for both LPs.

In general the models tended to perform better (in terms of classification) when the LPs were correlated. That may be due to the fact that there was less probability of certain combinations of levels of the learning progressions, so items tended to provide information regarding one ability estimate instead of having to parse out two different abilities. While Model 1 seemed to have high classification rates when the sample size was low and the sample used to generate the parameters was also used to classify students, this could be due to chance as the classification rates dropped when the sample size increased.

When a separate sample was used to estimate the classification then Model 1 and Model 2 tended to produce comparable results. When Model 1 and Model 2 were the models underlying the data generation then Models 3 and 4 also provided comparable results. However, when Model 3 was the generating model then Model 4 did not have as high a classification rate, and vice versa.

Conclusion

While parameter recovery for these models was fairly high the results of this study show that classification rates were not very high. These models did better when the two learning progressions were related to each other but still most of the time they classified more people incorrectly than correctly.

Of these models, Model 1 and Model 2 had the highest rate of classification overall. Even in the cases where these were not the generating model these models did a comparable job at classification. Model 2 may be appropriate in situations where it is important to keep the number of parameters low or there is a need to generate item IRT parameters along with the BIN. Model 1 may be used in other situations.

However, the overall results for this study demonstrated that the way the model was set up resulted in poor classifications and therefore a practitioner may not want to use a BIN for classifying students when they have only items are designed to measure two LPs. One possible reason for this is that there were no items that solely measured one LP. Therefore it may be the case that the LPs were reversed which would lead to low classification rates. Further studies are necessary to determine if there are ways to improve the classification rates of the BINs with multiple LPs.

CHAPTER 6: APPLICATION OF MODELS

This third study provides real-data applications in order to further explore the similarities and differences between the models in terms of fit and inferences. The data used is from a course offered through the Cisco Networking Academy. The Cisco Networking Academy is a global academy with several courses designed to help students obtain the knowledge and skills required for expertise in computer networking (Behrens et al, 2007).

This study is consists of two parts. Part A examines data focusing on one learning progression for IP addressing, while part B examines data that depends on two learning progressions, one for IP addressing and the other for Routing.

Part A: Real data that is designed to measure one LP

For part A the data set that was used was 36 items that depend on the IP addressing learning progression. While these items were not all from the same exam they were all taken on exams within the same month. The total sample size is 3827, which was partitioned into two subsamples. One sample of 1800 was used to estimate the conditional probabilities associated with the BIN and the remaining students were used as a cross-validation sample to test out the BIN. The learning progression for IP addressing has 5 levels (see Appendix A). However, for this study there were 4 items that are designed to provide evidence about whether a student is either below or at-or-above Level 1, 9 items that provide analogous evidence for Level 2, 12 items for Level 3, and 11 items for Level 4. As no items (as determined by content experts) were designed to measure the top level the LP was treated as only having 4 levels (with a 0 level as having no skill – i.e., below Level 1).

The overall process of this study was to first run a LCA to classify students into levels of the learning progression. Once students were classified, each of the four models from Study 1 (see Table 18 for a review of the models) was used to estimate parameters. Similarly to Study 1, fit statistics for each model were obtained and the estimated parameters were used to classify students both in the sample that generated the data and the separate sample. Results were then compared across models. (Note that unlike the simulations, this classification criterion is not a known true generating value, but rather an estimate from another, less constrained, model. Implications of this difference will be discussed in a subsequent section.)

The first decision that needed to be made was to determine how to assign subjects to levels of the learning progression. This study followed the approach used by West et al (2009) and used a latent class analysis. A five class model was fit which resulted in each student being assigned to a particular class, however, these classes, while labeled 1 through 5, did not necessarily correspond to levels of the learning progression. In order to determine the correct class, a mapping was made by examining the probabilities of a correct response to each of the items. By placing the probabilities in approximate order, the classes were able to be labeled with the group that had the lowest probabilities on items being labeled as Class 1 and the group with the highest probabilities being labeled as Class 5 (see Table 95 for the resulting classifications). Also note that while the probabilities for each individual item was not always in the same order (i.e. sometimes a higher class had a lower probability of a correct response than a lower class), the classifications tended to followed the content experts mapping of which items depend on which level by having the jump in probability be at the level for which the item was designed to provide evidence. For example, if an item was stated to be providing evidence on Level 1 then the probability for a student in Class 1 to obtain a correct response was relatively low compared to students with higher classifications. If an item was designed to provide evidence on Level 3 then the probabilities for students in Class 1 and Class 2 would be relatively lower than students in Class 3 or 4.

It should also be noted that the probabilities did not always fit into the expected pattern, in that some items had jumps in their probability at places other than where the content experts would have placed them. For example, item 18 has a high probability of a correct response for students in Class 2 but a low probability of a correct response for students in Class 3. This type of item is an example of an item that may not follow the hierarchical model and may in fact be indicative of the "messy middle" problem (as discussed in Chapter 3) and therefore Model 3 may be a more appropriate model than Model 1.

First BIN analysis

Once the assignment of classes was determined for each of the students the data was randomly split into two groups. One of these groups was used to determine the parameters for each of the four models discussed in Study 1 and the other group was used to examine the classifications. The data was then analyzed using the four models discussed in Study 1. All fit statistics indicated that Model 1 was the best fitting model (see Table 96). The percent of students correctly classified (i.e. classifications matched the LCA classification used as input when estimating probabilities in the BIN) reflected this as well (see Table 97).

Table 95: Probability of a correct response based on the latent class analysis. A change in shading reflects a jump in the probability The jumps were at least .05.

Level of the item	Item number	Class 1	Class 2	Class3	Class 4	Class 5
1	1	0.619	0.913	0.842	0.686	0.983
1	2	1.000	1.000	1.000	1.000	1.000
1	3	0.296	0.765	1.000	0.968	0.964
1	4	0.296	0.792	0.870	0.899	0.934
2	5	0.325	0.747	0.944	0.767	0.950
2	6	0.152	0.444	0.561	0.845	0.955
2	7	0.273	0.555	0.495	0.714	0.898
2	8	0.393	0.539	0.162	0.706	0.988
2	9	0.341	0.597	0.695	0.749	0.826
2	10	0.598	0.954	0.924	0.984	0.952
2	11	0.115	0.408	0.727	0.828	0.920
2	12	0.508	0.872	0.816	0.983	0.980
2	13	0.284	0.663	0.214	0.863	0.930
3	14	0.116	0.481	0.188	0.733	0.848
3	15	0.253	0.231	0.000	0.331	0.838
3	16	0.233	0.522	0.123	0.798	0.866
3	17	0.210	0.293	0.229	0.557	0.919
3	18	0.447	0.915	0.098	0.968	0.991
3	19	0.266	0.723	0.525	0.919	0.855
3	20	0.156	0.600	0.698	0.955	0.989
3	21	0.480	0.778	1.000	0.982	0.971
3	22	0.200	0.456	0.670	0.772	0.980
3	23	0.190	0.597	0.726	0.904	0.931
3	24	0.051	0.135	0.280	0.445	0.821
3	25	0.117	0.621	0.735	0.565	0.868
4	26	0.335	0.705	0.261	0.566	0.794
4	27	0.325	0.864	0.879	0.838	0.976
4	28	0.274	0.471	0.211	0.888	0.974
4	29	0.188	0.296	0.894	0.804	0.964
4	30	0.044	0.232	0.620	0.674	0.919
4	31	0.186	0.354	0.850	0.821	0.994
4	32	0.154	0.255	0.307	0.378	0.877
4	33	0.130	0.191	0.648	0.709	0.910
4	34	0.105	0.536	0.333	0.872	0.952
4	35	0.212	0.382	0.604	0.590	0.983
4	36	0.246	0.285	0.240	0.629	0.900

Table 96: Fit statistics for first BIN analysis of Part A of study 1: full data set

Fit Statistic	Model 1	Model 2	Model 3	Model 4
AIC	24219.3	33519.9	31248.5	31276.3
BIC	25230.5	33739.7	31484.8	31556.6
DIC	24172.0	35880.8	30968.9	30980.2

Table 97: Percent of students correctly classified for the first BIN analysis of Part A, study 1: full data set

Model	Percent classified Correctly	
	Parameter generating data	Separate data
1	69.7%	68.0%
2	30.7%	9.5%
3	30.2%	29.5%
4	28.9%	29.8%

Second BIN analysis

The percent classified consistently with the LCA by the first BIN analysis seemed low, especially for Models 2-4. Further examination of the data revealed that there was a large amount of missing data, and in fact a high percent of students had missing data for all items that were designed to provide evidence for particular levels of the learning progression. In those cases there was no direct information regarding whether or not the student has the particular attributes for that level. Due to this fact it was decided to remove any cases for which the student did not have responses to at least 2 items on every level. This resulted in a new sample size of 324. A latent class analysis was again run on this sample and class membership was computed in the same manner as before (see Table 98).

Table 98: Probability of a correct response based on the latent class analysis for the second BIN study. A change in shading reflects a jump in the probability

Level of the item	Item number	Class 1	Class 2	Class3	Class 4	Class 5
1	1	0.613	0.858	0.891	0.074	0.977
1	2	1.000	1.000	1.000	1.000	1.000
1	3	0.309	0.735	0.978	0.965	0.953
1	4	0.289	0.751	0.946	0.827	0.905
2	5	0.311	0.743	0.783	0.802	0.936
2	6	0.151	0.394	0.788	0.942	0.962
2	7	0.297	0.534	0.648	0.837	0.903
2	8	0.351	0.456	0.773	0.425	0.981
2	9	0.341	0.574	0.724	0.885	0.798
2	10	0.568	0.938	0.991	0.960	0.954
2	11	0.108	0.384	0.753	0.999	0.925
2	12	0.472	0.859	0.965	1.000	0.976
2	13	0.336	0.629	0.792	1.000	0.933
3	14	0.091	0.442	0.643	0.953	0.850
3	15	0.204	0.226	0.291	0.422	0.804
3	16	0.174	0.489	0.711	0.961	0.875
3	17	0.205	0.290	0.453	0.774	0.910
3	18	0.352	0.848	0.946	0.986	0.992
3	19	0.235	0.682	0.874	0.981	0.862
3	20	0.141	0.560	0.915	0.989	0.991
3	21	0.472	0.756	0.980	0.983	0.972
3	22	0.190	0.423	0.789	0.642	0.979
3	23	0.173	0.570	0.858	0.958	0.940
3	24	0.056	0.128	0.341	0.691	0.815
3	25	0.119	0.572	0.679	0.322	0.853
4	26	0.329	0.642	0.664	0.258	0.771
4	27	0.305	0.852	0.833	0.921	0.958
4	28	0.251	0.412	0.830	0.942	0.973
4	29	0.187	0.312	0.709	1.000	0.968
4	30	0.044	0.203	0.636	0.720	0.922
4	31	0.184	0.354	0.740	1.000	0.994
4	32	0.156	0.226	0.433	0.123	0.871
4	33	0.135	0.186	0.603	0.958	0.909
4	34	0.096	0.474	0.814	0.944	0.946
4	35	0.212	0.363	0.609	0.447	0.986
4	36	0.249	0.249	0.564	0.730	0.895

The probabilities of item responses were slightly different between when the full data set was used and when a subset of the data was used, but the general pattern of probabilities was very similar. One difference is that while in the full data set there were 11 items that had large drops in probabilities when moving into a higher level, with the small sample there were only 5 of these items.

The data was then split into two samples. A sample of 200 was used to generate the parameters and the other sample of 124 was used to cross validate the generated parameters. With the smaller sample size Model 3 was found to fit better (which follows along from some of the results that we saw in Study 1) (see Table 99). Again though, the percent of students whose classifications were consistent with the LCA was very low (see Table 100). This low classification rate may be due to the fact that the sample size was very small when compared to the number of items.

Table 99: Fit statistics for the second BIN analysis of Part A of study 1: Subset 1

Fit Statistic	Model 1	Model 2	Model 3	Model 4
AIC	7444.1	7283.1	7146.2	7173.1
BIC	8051.0	7415.0	7288.0	7341.3
DIC	7213.7	7240.5	7097.7	7108.7

Table 100: Percent of students correctly classified for the second BIN analysis of Part A of study 1: Subset 1

Model	Percent classified Correctly	
	Parameter generating data	Separate data
1	12.0%	33.1%
2	12.5%	18.5%
3	5.5%	19.4%
4	37.0%	17.7%

Third BIN Analysis

To test the theory that the sample size was not large enough for the number of items, a subset of items was picked from the total number of items. Three items from each level were picked such that the probabilities based on the latent class analysis had the largest jumps between the class that was one lower than the level the item was designed to measure and the class that was at the level the item was designed to measure. The items picked were 1, 3, 4, 6, 9, 10, 13, 14, 23, 29, 30 and 36.

With this subsample the fit statistics seemed to indicate that Model 1 was the best fitting model (see Table 101). This was supported with the classification rates (see Table 102). However, the classifications rates were still fairly low. The data was examined again and it was found that there were still some items with a fairly high number of students with missing data (for example item 1 for had missing data for 158 students while several other items had missing data for 197 students).

Table 101: Fit statistics for the third BIN analysis of Part A of study 1: Subset 2

Fit Statistic	Model 1	Model 2	Model 3	Model 4
AIC	3326.8	4069.1	3936.7	3948.0
BIC	3537.9	4161.5	4039.0	4050.3
DIC	2806.7	2926.6	2783.6	2794.9

Table 102: Percent of students correctly classified for the third BIN analysis of Part A of study 1: Subset 2

Model	Percent classified Correctly	
	Parameter generating data	Separate data
1	46.0%	26.0%
2	7.5%	25.8%
3	25.0%	22.6%
4	24.5%	23.4%

Fourth Bin Analysis

For the fourth BIN analysis a different subset of items was chosen in which the highest number of missing data points was 14. These items were 2, 3, 4, 5, 10, 11, 19, 22, 24, 30, 35, 36. The data was again split into two samples with 200 in the sample for generating the parameters and 124 in the sample for testing out these parameters. The AIC and BIC fit statistics indicated Model 1 was the best fitting model. The DIC indicated Model 3 (see Table 103). The classification rates also indicated that Model 1 was the best fitting model as it had the highest classification rate (about 20% higher than the other models) (see Table 104). However, even Model 1 misclassified more students then classified correctly.

Table 103: Fit statistics for the fourth BIN analysis of Part A of study 1: Subset 3

Fit Statistic	Model 1	Model 2	Model 3	Model 4
AIC	3201.6	4165.7	4019.3	4036.6
BIC	3412.7	4258.0	4121.5	4138.8
DIC	3054.8	3157.4	3014.9	3026.1

Table 104: Percent of students correctly classified for the fourth BIN analysis of Part A of study 1: Subset 3

Model	Percent classified Correctly	
	Parameter generating data	Separate data
1	42.0%	44.4%
2	11.0%	29.8%
3	22.5%	24.2%
4	22.0%	23.4%

Discussion of Part A

The results of this study indicate that of these models Model 1 performs the best. However further studies are needed to justify the use of a BIN. One factor in this study was the amount of missing data. Further studies may want to examine how much missing data is acceptable before the model does not perform well in terms of classification. Studies may want to examine missing data when generating the parameters and/or missing data when trying to determine the correct classification of students.

Another issue with this study may be from the assignment of students to classes. For this study the only information that was available regarding each of the students was the current data set. Therefore it is not guaranteed that the assignment to levels of the learning progression is correct, in which case trying to determine how well the models matched the original classification may not provide the most accurate information on how well students would be correctly classified. Also, while this study assumed that there was an underlying hierarchical structure as described by the content experts, some of the items did not follow this pattern as can be seen by the decrease in probability of correct response as the level increased. Results might have been different if classifications for the individual levels were available. Further studies may want to find other sources of information to use to determine correct levels for each student in the sample used to learn parameters of the models. In addition it may be interesting to determine how well the LCA is able to classify students, as part of this study is a comparison between two models that can be used for classification and it is not clear that the classifications from the BIN would be any better than the classifications from an LCA.

Overall, while Model 1 and Model 3 seemed acceptable based on the simulation study, the real data example demonstrates that there are situations in which neither of these models would provide acceptable classification rates. Further information into how these models would behave in real-data situations may be desired.

Part B: Real Data Study with Two LPs

For part B a data set was used that took items from a final exam in the Cisco Networking Academy. The data set included 6 items designed to measure students ability with regards to IP addressing, 6 items that were designed to measure students skills in routing and 10 items that were designed to measure both. For this data set there was no missing data and the total sample size was 831.

Each LP had 5 levels (see Appendix A for discussion of the levels). The levels that each of the items was designed to provide evidence on can be seen in Table 105. The number of items that were designed to measure each level of the learning progression was determined (see Table 106). Since there were no items at level 1 of the IP addressing LP and only 2 items at level 5, a LP with 3 levels (plus a novice level) was used. These three levels would correspond to level 2 of the original LP, level 3 of the original LP, and levels 4 and 5 of the original LP. Similarly for the Routing IP 3 levels were used, level 1 corresponding to levels 1 and 2 of the original LP, level 2 corresponding to level 3 of the original LP, and level 3 corresponding to level 4 of the original LP.

Table 105: The levels of each LP that the items are designed to measure

Item Number	Level of IP Add. LP	Level of Router LP
1	4	3
2	4	2
3	5	3
4	3	3
5	2	2
6	2	3
7	3	3
8	3	2
9	2	2
10	2	3
1	4	-
2	4	-
3	5	-
4	3	-
5	3	-
6	3	-
1	-	1
2	-	3
3	-	4
4	-	4
5	-	4
6	-	4

Table 106: The number of items designed to measure each level of the LPs

Level of the LP	Items on IP Add.	Items on Routing
1	0	1
2	4	4
3	6	7
4	4	4
5	2	0

Two latent class analyses were used to determine how to assign levels to each student. The first analysis used the items that were designed to provide evidence on IP addressing and the items measuring both LPs and the second analysis used the items that were designed to provide evidence on routing and both LPs. The items that were designed to measure both LPs were used, as the items measuring just one learning progression did not always provide information on all levels of the learning progression and so it was deemed that these items would not provide enough information for classification purposes. Using these items was not ideal, as this part of the study was trying to classify students on the individual LPs but the response probabilities for these items are based on student's ability on both the LPs. A better approach would have been to determine the level of the students outside of this assessment; however this type of information was not available.

Once the latent class analysis was run the resulting classes were arranged in order to best reflect the content experts mapping of the items to the level of the learning progression that it is aimed at measuring and to order the probabilities so that lower classes had lower probability of responses (see Table 107 and Table 108). For both of these analyses the items that only depended on one of the LPs were arranged so that the probabilities increased as the levels increased. In this arrangement the probabilities for items measuring both of the LPs did not always increase as the level increased. This could be in part due to the fact that these items also depended on another skill, or it could be part due to the fact that students might not obtain all of the level attributes in order.

Table 107: Probability associated with each class based on the latent class analysis for IP addressing

Items Measuring	Item Number	Level of IP Add. LP	Class 1	Class 2	Class 3	Class 4
Both LPs	5	2	0.365	0.854	0.740	0.834
	6	2	0.234	0.076	0.493	0.410
	9	2	0.337	0.516	0.735	0.715
	10	2	0.354	0.468	0.444	0.694
	4	3	0.210	0.904	0.693	0.922
	7	3	0.321	0.261	0.423	0.320
	8	3	0.385	0.405	0.685	0.829
	1	4	0.198	0.910	0.503	0.968
	2	4	0.091	0.765	0.180	0.823
	3	5	0.054	0.900	0.323	0.937
IP Add.	4	3	0.450	0.374	0.856	0.757
	5	3	0.194	0.107	0.297	0.411
	6	3	0.422	0.408	0.670	0.736
	1	4	0.388	0.422	0.463	0.864
	2	4	0.314	0.308	0.753	0.798
	3	5	0.414	0.367	0.843	0.785

Table 108: Probability associated with each class based on the latent class analysis for Routing

Items Measuring	Item Number	Level of Routing.	Class 1	Class 2	Class 3	Class 4
Both LPs	2	2	0.365	0.854	0.740	0.834
	5	2	0.234	0.076	0.493	0.410
	8	2	0.337	0.516	0.735	0.715
	9	2	0.354	0.468	0.444	0.694
	1	3	0.210	0.904	0.693	0.922
	3	3	0.321	0.261	0.423	0.320
	4	3	0.385	0.405	0.685	0.829
	6	3	0.198	0.910	0.503	0.968
	7	3	0.091	0.765	0.180	0.823
	10	3	0.054	0.900	0.323	0.937
Routing.	1	1	0.450	0.374	0.856	0.757
	2	3	0.194	0.107	0.297	0.411
	3	4	0.422	0.408	0.670	0.736
	4	4	0.388	0.422	0.463	0.864
	5	4	0.314	0.308	0.753	0.798
	6	4	0.414	0.367	0.843	0.785

The data was then split into two samples, a sample size of 400 was used to generate parameters and a sample size of 431 was used to test the classification results. The four models described in Study 2 were each applied (see Table 41 for description of the models). The AIC and DIC indicated that Model 1 was the best fitting model (see Table 109). The BIC indicated that Model 4 was the best fitting model.

Table 109: Fit statistics for each model in Part B of study 3

Fit Statistic	Model 1	Model 2	Model 3	Model 4
AIC	11579.8	12201.7	12134.2	12127.5
BIC	12434.0	12349.4	12281.9	12275.2
DIC	11271.6	12161.6	12094.6	12088.0

Classification results from first set of items

The percent of students whose classification matched the classification generated from the LCA was calculated for both the sample that generated the parameter and the separate sample (see Table 110). In all cases, the classification rates for Model 1 were at least 15% higher than all of the other models. Models 2-4 had classification rates in the range 45.5% to 68.2% which was higher than those found in the simulation study. One possible explanation for this is that in this study there were items that were only dependent upon the individual LPs The adjusted Rand statistic followed the same pattern, with Model 1 having higher rates and in general the rates were higher than in the simulation study (see Table 111).

Table 110: Classification rates for Study 3, Part B when all items were used

	Generating Parameters Sample			Separate Sample		
Model	LP1	LP2	Both	LP1	LP2	Both
1	89.5%	90.5%	82.0%	88.4%	83.5%	74.0%
2	68.0%	65.8%	54.0%	64.5%	66.1%	51.3%
3	66.5%	56.3%	45.5%	67.1%	59.4%	47.8%
4	64.5%	68.0%	48.5%	68.2%	68.2%	49.4%

Table 111: The adjusted Rand index Study 3, Part B when all items were used

	Generating Parameters Sample			Separate Sample		
Model	LP1	LP2	Both	LP1	LP2	Both
1	0.741	0.767	0.782	0.727	0.614	0.680
2	0.418	0.396	0.464	0.390	0.373	0.423
3	0.493	0.302	0.459	0.513	0.353	0.474
4	0.431	0.442	0.494	0.461	0.406	0.449

Classification results from using a subset of items

To determine if having the items that measured both LPs increased classification rates over only using items that measured one LP the classification rate was calculated for the samples using only the information from the items that measure one LP (see Table 112). The classification rates for Models 3 and 4 decreased slightly; about a 2% decrease on the classifications for both LPs. Models 1 and 2 had much higher decreases of around 20%-30%. Overall the classification rates were more similar to each other than they were when all the items were used, which makes sense as the only difference between the models is how they handle the situation when items depend on multiple LPs and items that only depend on one LP would have the same structure across the models. It also implies that when there are items that depend on multiple LPs, Model 3 and Model 4 may not be appropriate as the use of these items do not improve the classification of students for these models. Model 1 had the highest gain in classification rate which may indicate that practitioners would want to use Model 1.

Table 112: Classification rates for Study 3, Part B when only items measuring one LP were used

	Generating Parameters Sample			Separate Sample		
Model	LP1	LP2	Both	LP1	LP2	Both
1	61.8%	59.8%	44.3%	64.0%	58.0%	45.0%
2	54.8%	51.8%	40.5%	51.3%	54.3%	39.0%
3	54.8%	56.5%	44.3%	55.0%	59.4%	45.7%
4	57.5%	55.0%	46.3%	58.9%	58.5%	47.3%

The adjusted Rand indices had a decrease for all of the models, with Model 1 having the largest decrease (see Table 113). The statistics were also very similar across models, although Model 1 did have the highest statistic (except for the when matching the classification on LP2 for the separate sample in which case it was .001 below that of Model 4). This indicates that having the items that depend on two LPs increased the match between the resulting classification and the starting classification. This increase was largest for Model 1 which again may indicate that when having items that measure multiple LPs, Model 1 may be the model that would provide the highest classification rate.

Table 113: The adjusted Rand index Study 3, Part B when only items measuring one LP was used

	Generating Parameters Sample			Separate Sample		
Model	LP1	LP2	Both	LP1	LP2	Both
1	0.318	0.341	0.389	0.379	0.327	0.403
2	0.281	0.301	0.359	0.308	0.346	0.364
3	0.283	0.296	0.343	0.296	0.353	0.345
4	0.281	0.285	0.343	0.307	0.328	0.353

Discussion of Part B

While the results from study 2 did not show promise in the use of a BIN when measuring multiple LPs, the results of this study demonstrate that there may be situations for which this type of model would be appropriate. Of the four models used Model 1, which did not constrain the relationship between the two LPs, had much higher classification rates than the constrained models when all of the data was used.

As with the previous example there is some concern regarding how students were assigned to groups. The only information provided was the current data set, which had limited coverage in the items that were designed to measure one skill. In follow up studies it may be useful to have other information to be used to classify students such as teacher ratings and previous coursework.

One highlight of this study is that is seems that the use of items that measure just one LP along with items that measure both LPs can help the accuracy of the classification. Follow up studies may want to examine this issue to determine if there is an optimal or minimum number of items for measuring one skill and items measuring multiple skills.

Conclusion

Overall this study highlighted some of the issues with using a BIN in practice. The first issue is the determination of the classification of students for the sample that will be used to generate the parameters. Ideally information outside of the results of the exam would be used such as teacher input and results from previous assessments. The accuracy of the BIN depends in part on the accuracy of these classifications.

Another highlight from this study is the issue of missing data. Part A demonstrated that having a large amount of missing data can decrease the classification rates for a BIN. Further studies may be needed to determine how much missing data and the types of missing data that would most affect the situation, but overall practitioners may want to be careful if they are using a data set with a large amount of missing data to estimate the BIN parameters.

Part B demonstrated that using a combination of variables that are designed to measure one LP and variables that are designed to measure multiple LPs can help improve classification rates over just variables that are designed to measure multiple LPs. Practitioners may want to incorporate both types of variables when they are designing their assessment.

CHAPTER 7: CONCLUSIONS

Summary of Findings

This research was designed to address different modeling issues with regards to using BINs to model LPs. The work included three studies, the first to examine modeling issues when each observable variable is designed to measure levels on one learning progression, the second to address issues when the observable variables measure two learning progressions, and the third study to provide real data examples of how these techniques can be applied in practice.

The goal of the research was to provide insight into the cases in which different models may be more appropriate in practical applications. When dealing with one learning progression, four models were compared. The first model treated the LP as one latent categorical variable, while the other models treated the LP as having separate binary variables for each level. In this latter case, three models were compared, one in which the binary variables were independent of each other, one in which each variable was dependent on the variable associated with the previous level of the LP, and the last in which each latent variable was dependent on the variables associated with all of the previous levels of the LP.

The simulation study indicated that treating the LP as one categorical latent variable was the only model in which the classification rates were higher than 75% in all cases when the generating sample was used for classification rates, and 65% when a separate sample was used. Treating the LP as independent level variables generally resulted in the lowest classification rates of all models. Treating the LP as different level variables with a dependency between adjacent levels, seemed to produce comparable or higher classification rates, compared to treating the LP as one categorical latent variable under two conditions: when the sample size was small, or the ability distribution of students was skewed.   The real data example, while not providing high classification rates (where the classifications were based on LCA results, so may not be the correct classifications) did indicate that treating the LP as one categorical latent variable may provide the highest classification rates among the models being compared.

The issue of whether or not there was a benefit in adding in constraints regarding the structure of the dependence relationship of an OV that is modeled as depending on two LPs, was addressed in Study 2. The benefits of adding in constraints would be that there would be fewer parameters to estimate and that constraints could be made such that items could be placed on a familiar IRT scale to represent their difficulty. Study 2 compared an unconstrained model with three constrained models that used compensatory, conjunctive and disjunctive relationships.

The results of the simulation study indicate that while an unconstrained model or a compensatory model would provide comparable or higher classification rates than the conjunctive or disjunctive model, neither the constrained nor the unconstrained model were able to consistently classify students correctly over 50% of the time.   The conjunctive and disjunctive models were not as robust to model misspecification and therefore would not be recommended, at least not at the sample sizes used in this study. The real data example seemed to indicate that the addition of variables that measured only one of the LPs might increase the classification rates of the models particularly for the unconstrained model. However, further studies would be needed to validate this indication.

Limitations

This study is only a beginning of investigations into applying BINS to LPs. One limitation with Study 1 and Study 2 is that a couple of the assumptions about the models may not reflect real world situations. For one, the relationship between the OV and LPs was fixed to specify a strong relationship of the OVs with the level they were designed to measure. Also, the probability of a correct response was the same (and low) for levels below where an item was designed to measure, then jumped to the higher value and did not change. These assumptions may not be true in a real world example as items may display a more gradual change in probability across levels. Additionally, in these studies there were multiple indicators of each level of the learning progression. As can be seen from the real data example, this might not always be the case.

The real data examples in this study also had some limitations. The main issue was the method in which the students were assigned to classes. For this study the only information that was given with regards to the student was the test scores, and therefore the classifications were derived from the same data that was used for the analysis. One issue with this approach, particularly in the first example, was that this method attempted to impose a hierarchical structure on the LP which might have biased the analysis towards the model that closely matched that structure. Also, in the real data example it was not known if the classifications assigned were correct. The study was then comparing classifications from the BIN to classifications that may or may not be correct, and the poor results may be due to differences in the methods versus an issue with a BIN.

The missing data in Part A of the real data example was an additional limitation. While this data set provided insight into how missing data might affect the results, it did not provide an opportunity to examine a data set that was close to the situation found in the simulation study.

Another limitation with Part B of the real data example is that there were not enough items on every level of the learning progression, in particular on level 1. The study was not able to separate students who were at level 1 from students who were at level 2 for either of the LPs.

Implications for Practice

A practitioner who is using a BIN for measuring an LP may want to decide how important is it obtain information on the individual levels of the learning progression versus the learning progression as a whole. In addition a practitioner may want to determine if they believe the sample they are using has a skewed ability distribution. If that is the case then they may want to split up the learning progression into individual level variables. Otherwise they can just use one variable to represent the entire LP.

When designing an assessment that is used to measure two LPs, a practitioner would want to include items that solely measure each individual skill in addition to items that may measure both. In this situation an unconstrained model is recommended. However, a practitioner may want to be cautious about using the results of a BIN for high stakes situations as there may be a high level of misclassification.

In addition, a practitioner needs to decide how they are going to gather data to learn the parameters of the BIN. They will want to use a data set that already has classification information associated with each student – that is, “supervised learning” rather than “unsupervised learning” for the conditional probabilities in the BIN. In this case, the decision must be made as to how this gold-standard information will be determined. (One example may be from teacher judgments based on extensive information from classroom and hands-on performances.)

Directions for Future Research

This research can be extended in several ways, such as examining the case where the data is polytomous, including different types of relationships between the latent variables and the observables, examining how the number of items used affects classification rates, and examining more complex learning progressions. In particular it would be useful to examine the situation in which two LPs are being measured and to vary the number of items that are measuring each LP as well as measuring both LPs.

Another area of interest would be to compare the performance of BINs with other models adapted to be used with LPs. One possibility is instead of breaking out the attributes into separate LP level variables, breaking them out into different attributes as input to a CDM. Another possibility is to apply an IRT model and to set cutpoints that would separate out the levels of the LP. In particular since this study did not indicate that, within the conditions studied here, BIN would be a good model to use in the case where multiple LPs are being measured one could investigate whether there are other models that would do better, or if the less-than-satisfactory performance should be attributed to the carrying capacity of the data; that is, whether given the data, any model can recover underlying progress levels. It may be the case that the numbers and the structures of tasks needed to assess students’ levels on learning progressions are larger than are needed for more familiar kinds of overall proficiency assessments.

This study provided some insight into when particular BINs would be appropriate (or not appropriate). However there is still room for more research in this field and the field of modeling LPs. As LPs become more popular there will be a greater need to develop models that provide the accuracy required to provide useful information from assessments designed to measure LPs. Further work can be done in order to provide practitioners with insight that will help them develop and model assessments for LPs.

APPENDIX A

A1: IP Addressing Learning Progression

Level	Knowledge and skills
1: Novice	1. Student can navigate the operating system to get to the appropriate screen to configure the address.
	2. Student knows that four things need to be configured: IP address, subnet mask, default gateway and DNS server.
	3. Student can enter and save information.
	4. Student can use a web browser to test whether or not network is working.
	5. Student can verify that the correct information was entered and correct any errors.
	6. Student knows that DNS translates names to IP addresses.
	7. Student understands why a DNS server IP address must be configured.
2: Basic	1. Student understands that an IP address corresponds to a source or destination host on the network.
	2. Student understands that an IP address has two parts, one indicating the individual unique host and one indicating the network that the host resides on.
	3. Student understands how the subnet mask indicates the network and host portions of the address.
	4. Student understands the concept of local –vs- remote networks.
	5. Student understands the purpose of a default gateway and why it must be specified.
	6. Student knows that IP address information can be assigned dynamically.
	7. Student can explain the difference between a broadcast traffic pattern and a unicast traffic pattern.
3: Intermediate	1. Student understands the difference between physical and logical connectivity.
	2. Student can explain the process of encapsulation.
	3. Student understands the difference between Layer 2 and Layer 3 networks and addressing.
	4. Student understands that a local IP network corresponds to a local IP broadcast domain. (both the terms and the functionality)
	5. Student knows how a device uses the subnet mask to determine which addresses are on the local Layer 3 broadcast domain and which addresses are not.
	6. Student understands the concept of subnets and how the subnet mask determines the network address.
	7. Student understands why the default gateway IP address must be on the same local broadcast domain as the host.
	8. Student understands ARP and how Layer 3 to Layer 2 address translation is accomplished.
	9. Student knows how to interpret a network diagram in order to determine the local and remote networks.
	10. Student understands how DHCP dynamically assigns IP addresses.

4: Advanced	1. Student can use the subnet mask to determine what other devices are on the same local network as the configured host.
	2. Student can use a network diagram to find the local network where the configured host is located.
	3. Student can use a network diagram to find the other networks attached to the local default gateway.
	4. Student can use the PING utility to test connectivity to the gateway and to remote devices.
	5. Student can recognize the symptoms that occur when the IP address or subnet mask is incorrect.
	6. Student can recognize the symptoms that occur if an incorrect default gateway is configured.
	7. Student can recognize the symptoms that occur if an incorrect DNS server (or no DNS server) is specified.
	8. Student knows why DNS affects the operation of other applications and protocols, like email or file sharing.
	9. Student can use NSlookup output to determine if DNS is functioning correctly.
	10. Student can configure a DHCP pool to give out a range of IP addresses.
	11. Student knows the purpose of private and public IP address spaces and when to use either one.
	12. Student understands what NAT is and why it is needed.
5: Expert	1. Student can recognize a non-functional configuration by just looking at the configuration information, no testing of functionality required.
	2. Student can interpret a network diagram to determine an appropriate IP address/subnet mask/default gateway for a host device.
	3. Student can recognize the symptoms that occur if an incorrect subnet mask is configured on the intermediate routers or destination host.
	4. Student can interpret a network diagram in order to determine the best router to use as a default gateway when more than one router is on the local network.
	5. Student can evaluate a connectivity problem to determine if it could possibly be caused by an incorrect setting configured on the host.
	6. Student can propose changes to a host configuration to solve a connectivity problem.
	7. Student can make and test proposed changes to a host configuration to solve an identified connectivity problem.
	8. Student can implement NAT to translate private to public addresses.

A2: Routing Learning Progression

Level	Knowledge and skills
1: Novice	1. Differentiate Layer 2 networks from Layer 3 networks.
	2. Understand the difference between local and remote networks
	3. Understand the relationship of IP network address to local physical network.
	4.Understand how a host uses its own subnet mask to determine if a destination address is on the same local network
	5.Explain network broadcast messages and their purpose in a network.
	6. Understand that ARP messages do not leave the local Layer 3 network.
	7. Understand that the function of a gateway is to forward packets from one network to another.
	8. Understand that the routing process is required to get packets from the source local network to the destination network.
	9. Understand that routers use network layer addresses to get packets from the source local network to the destination network.
	10. Interpret a network diagram to determine when routing is necessary for a packet to be sent from one host to another.
2: Basic	1. Realize that routing is a function, not a device, and that any computer with two NICs can perform the routing function.
	2. Understand that routers do not normally forward broadcasts from one network to another, so routers form the boundary of a broadcast network.
	3. Realize that transmission media and Layer 2 protocol can change from one router interface to another.
	4. Explain the differences between LAN and WAN.
	5. Understand that routers remove the frame headers and re-frame the packet for transmission.
	6. Realize that a router LAN interface is another host on the local network and operates in many of the same ways that other hosts do. (responds to ARPs, originates and respond to PINGs, processes broadcasts, has MAC address.
	7. Understand that a routing device may also perform other functions, such as running management, client/server, and configuration software
	8. Differentiate between directly connected, static, and default routes.
	9. Understand how routers keep tables containing destination networks and the router interfaces to use to reach them.
	10. Explain classful networking and how some functions still rely on network classes (example: default subnet mask, "network x is subnetted" output in a routing table)
	11. Perform a basic router configuration.
	12. Use show commands to display router configurations and the contents of the routing table
	13. Interpret a routing table that contains directly connected, static, and default routes.
	14. Configure simple static and default routes.
	15. Understand the relationship between the status of an interface and the contents of the routing table

	16. Use the "show interface" command output to determine the status of an interface.
3: Intermediate	1. Understand the concept of segmented networks and the meaning of the term "hops".
	2. Explain the benefits of network segmentation.
	3. Understand the concept of a point-to-point network and why connections between routers are often point-to-point
	4. Understand the role of the subnet mask in the destination network path selection process.
	5. Explain the concept of "longest match".
	6. Know that routing protocols enable routers to exchange Layer 3 information.
	7. Know that routers use broadcasts and multicasts to exchange information
	8. Know that Cisco Discovery Protocol is not a Layer 3 routing protocol, that it uses a Layer 2 frame to enable the exchange of device specific information between directly connected Cisco devices
	9. Explain the advantages of statically configured and dynamically learned routes, including the fact that static routes take priority over dynamically learned routes
	10. Interpret a network diagram in order to select the appropriate default route.
	11. Explain the concept of route metrics, using distance vector examples
	12. Interpret a routing table to determine which route will be used for any destination address
	13. Configure a dynamic routing protocol (RIPv2) to advertise directly connected routes
	14. Verify the operation of a dynamic routing protocol (RIPv2) using show commands
	15. Troubleshoot routing problems related to network statement configuration errors
4: Advanced	1. Describe the differences between different routing protocols. (IGP/EGP,distance vector/link state,classful/classless,EIGRP, OSPF,RIPv1/v2)
	2. Understand how RIPv2, EIGRP and OSPF exchange information and select routes
	3. Understand the concept of neighbor routers and the various roles routers may perform in a complex network
	4. Interpret a network diagram to determine how a specific routing protocol will select the best route to a destination (example: given this diagram, OSPF will use this route...)
	5. Explain why some routing protocols require other tables to be stored on the router (topology, neighbor, successor, etc.)
	6. Understands the concept of administrative distance and how it can be manipulated and verified to ensure a specific route is installed in the routing table
	7. Understands that multiple routing protocols can be active on a router at the same time and that information learned using one method can be redistributed (shared) through another

	8. Understands the concept of route summarization and the importance of a hierarchical addressing structure
	9. Understand when static routing is preferable to dynamic routing and why
	10. Configure a combination of static and dynamic routing using RIPv2, EIGRP or single area OSPF
	11. Use show and debug commands to determine if routing information is being correctly sent and received
5: Expert	1. Understand the importance of authenticating routing protocol neighbors in order to trust the routing updates
	2. Understand how routing loops can cause network instability and the mechanisms that routing protocols use to prevent them
	3. Explain how floating static routes work and when they should be used
	4. Understand the difference between how Interior Gateway Protocols exchange information and how Exterior Gateway Protocols exchange information
	5. Understand the concept of network area borders and the function of a border router
	6. Interpret a network diagram to determine which routing method will best meet needs
	7. Predict which routes will be installed in a routing table given a network diagram and show run output from network routers
	8. Configure optimal route summarization
	9. Configure a routing protocol to appropriately redistribute static and default routes
	10. Adjust features of routing protocols to suit communication needs.
	11. Troubleshoot common issues with RIPv2, EIGRP, and OSPF

REFERENCES

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J. D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44,* 341-359.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*.

Almond, R. G., Mulder, J., Hemat, L.A., & Yan, D. (2009). Bayesian network models for local dependence among observable outcome variables. *Journal of Educational and Behavioral Statistics, 34*(4), 491-521.

Almond, R. G., Yan, D., & Hemat, L. (2008). Parameter recovery studies with a diagnostic Bayesian network model. *Behaviormetrika*, *35*(2), 159-185.

Behrens, John T., Frezzo, D. C., Mislevy, R. J., Kroopnick, M., & Wise, D. (2007). Structural, functional, and semiotic symmetries in simulation-based games and assessments. In E. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.) *Assessment of problem solving using simulations* (pp. 59-80). New York: Earlbaum

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

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, pg 261.

Chi, M. T. H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5,* 121-152.

Corcoran, T., Mosher, F., Rogat, A. (2009) *Learning progression in science: An evidence-based approach to reform.* Philadelphia, Pa: Consortium for Policy Research in Education.

Corrigan, S., Loper, S., Barber, J., Brown, N., & Kulikowich, J. (2009). *The Juncture of supply and demand for information: How and when can learning progressions meet the information demands of curriculum developers?* Paper presented at the Learning Progressions in Science (LeaPS) Conference. Iowa City, IA.

Cowles, M.K., & Carlin, B. P. (1995). Markov chain Monte Carlo diagnostics: A comparative review. *Journal of the American Statistical Society, 91*, pp 883-904.

Dayton, C. M. & Macready, G. B. (1980). A scaling model with response errors and intrinsically unscalable respondents. *Psychmetrika, 45*(3), pp343-356.

De la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*.

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

DeBarger, A. H., Ayala, C., Minstrell, J., Kraus, P., & Stanford., T. (2009). Facet based progressions of student understanding in chemistry. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Decision Systems Laboratory, University of Pittsburgh. (2003). *Genie reference manual*. Retrieved Dec 10, 2011 from: http://genie.sis.pitt.edu/wiki/GeNIe_Documentation

Dibello, L. V., Roussos, L.A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C.R. Rao & S. Sinharay

(Eds.), *Handbook of Statistics* (pp. 979-1030) Vo. 26, Psychometrics. , Elsevier Science B.V.: The Netherlands.

Draney, K. (2009). *Designing learning progressions with the Bear assessment system*. Paper presented at the Learning Progressions in Science (LeaPS) Conference. Iowa City, IA.

Drasgow, F., Dorans, N. J., & Tucker, L. R. (1979). Estimators of the squared cross-validity coefficient: A Monte Carlo investigation. *Applied Psychological Measurement, 3*(3), 387-399.

Ericsson, A.K., Charness, N., Feltovich, P., & Hoffman, R.R. (2006). *Cambridge handbook on expertise and expert performance*. Cambridge, UK: Cambridge University Press.

Formann, A.K. (2003). Latent class model diagnosis from a frequentist point of view. *Biometrics, 59,* pp 189-196.

Formann, A.K. & Kohlmann, T. (1998). Structural latent class models. *Sociological Methods & Research, 26*(4), 530-565.

Gagne, R. M., (1970). Basic studies of learning hierarchies in school subjects. Final report. Report for the Office of Education, Washington, D. C. Bureau of Research. Downloaded July, 2010 from http://www.eric.ed.gov/PDFS/ED039611.pdf.

Gagne, R. M & Driscoll, M. P. (1988). *Essentials of learning For instruction, 2nd edition.* Englewood Cliffs, NJ: Prentice Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B (2004). *Bayesian data analysis Second Edition*. Boca Raton, FL: Chapman & Hall/CRC.

Gilks, W. R., Richardson, S., & Spiegelhalter, D.J. (1996). *Markov chain Monte Carlo in Practice*. Boca Raton, FL: Chapman & Hall/CRC.

Gotwals, A. W., Songer, N. B. (2009). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Wiley InterScience*. Retrieved Oct 2011 from www.interscience.wiley.com.

Gotwals, A. W., Songer, N. B., & Bullard, L. (2009). *Assessing student's progressing abilities to construct scientific explanations.* Paper presented at the Learning Progressions in Science (LeaPS) Conference. Iowa City, IA.

Gunckel, K. L., Covitt, B. A., & Anderson, C. W. (2009). *Learning a secondary discourse: Shifts from force-dynamic to model-based reasoning in understanding water in socio-ecological systems.* Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Haberman, S. H., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75(2),* pp209-227.

Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: implications for testing. In N. Frederiksen, R. J., Mislevy and I. I Bejar (Eds.), *Test theory for a new generation of tests* (pp 359-384). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston, MA: Kluwer Nijhoff Publishing.

Huff, K. & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. Leighton and M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press.

Koski, T. & Noble, J. M. (2009). *Bayesian networks an introduction*. Great Britain: John Wiley & Sons, Ltd.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research, 33, 188-229.*

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

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

Levy, R., Crawford, A. V., Fay, D., & Poole, K. L. (2011). Data-model fit assessment for Bayesian networks for simulation-based assessments. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Levy, R. & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing, 4*(4), 333-369.

Li, F., Cohen, A. S., Kim, S., Cho, S. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*.

Liu, C. (2009). Selecting Bayesian-network models based on simulated expectation. *Behaviormetrika, 36*(1), pp. 1-25.

Liu, Y., Douglas, J. A., Henson, R. A (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement, 33*.

Liu. X. (1992, June). *The Dimensionality of test data generated by compensatory and non-compensatory two-dimensional IRT models and its effect on model-data-fit*. Paper presented at the Annual Meeting of the Canadian Society for the Study of Education, Charlottetown. Prince Edward Island.

McCutcheon, A. (1987). *Latent class analysis*. Newbury Park, CA: Sage Publications.

Mignami, S. & Rosa, R. (2001). Markov chain Monte Carlo in statistical mechanics: The problem of accuracy. *Technometrics*, 43(3), pp 347-355.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds). *Test theory for a new generation of tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychmetrika, 59*(4), 439-483.

Mislevy, R.J. (1997). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment*, pp. 43-71. Hillsdale, NJ: Erlbaum.

Mislevy, R. J., Almond, R., Dibello, L., Jenkins, F. Steinberg, L., and Yan, D. (2002). Modeling conditional probabilities in complex educational assessments. *CSE Technical Report 580.* Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Mislevy, R. J., Almond, R. G., and Lukas, J. (2003). A brief introduction to evidence centered assessment design. *Research Report RR-03-16.* Princeton, NJ:

Educational Testing service. Retrieved October 31, 2011, from http://www.ets.org/Media/Research/pdf/RR-03-16.pdf .

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Lawrence Erlbaum.

Mohan, L., & Anderson, C. W. (2009). *Teaching experiments and the carbon cycle learning progression*. Paper presented at the Learning Progressions in Science (LeaPS) Conference. Iowa City, IA.

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

Nichols, P. D, Chipman, S. F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic asssessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Norsys Software Corporation. (2007). *Netica manual.* Retrieved June 15, 2010 from: http://www.norsys.com/netica.html.

Piaget, J. (1928). Psychopédagogie et mentalité enfantine. *Journal de psychologie normale et pathologique* (Paris), 25, 31-60.

Popham, W. J. (2007). The lowdown on learning progressions. *Educational Leadership*.

R Development Core Team (2008*). R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http//www.R-project.org.

Reckase, M. D. (2009). *Multidimensional item response theory*. London: Springer.

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

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

Schum., D. A. (1994). *The evidential foundations of probabilistic reasoning*. New York: Wiley.

Schwarz, C., Reiser, B., Fortus, D., Shwartz, Y., Acher, A., Davis, B., Kenyon, L., & Hug, B. (2009). *Models: Defining a learning progression for scientific modeling*. Paper presented at the Learning Progressions in Science (LeaPS) Conference. Iowa City, IA

Sinharay, S. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics, 29*, pp 461.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics, 31*.

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

Spiegelhalter, D. J., Best, N. G., & Carlin, B. P.(1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Research Report 98-009. Division of Biostatistics, University of Minnesota.

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

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). WinBugs version 1.4: User manual. Cambridge Medical Research Council Biostatistical Unit. http://www.mrc-bsu.carn.ac.uk/bugs/.

Stevens, S. Y., Shin, N., & Krajcik. (2009). *Towards a model for the development of an empirically tested learning progression*. Paper presented at the Learning Progressions in Science (LeaPS) Conference. Iowa City, IA.

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods. 9*:3.

Templin, J. L., & Henson, R. A (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305.

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

West. P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., DiCerbo, K. E., Crawford, A., Choi, Y., & Behrens, J. T. (2009). *A Bayes net approach to modeling learning progressions and task performances.* Paper presented at the Learning Progressions in Science (LeaPS) Conference. Iowa City, IA.

Wheeler, D. C., Hickson, D. A., & Waller, L. A. (2010). Assessing local model adequacy in Bayesian hierarchical models using the partitioned deviance information criterion. *Computational Statistics and Data Analysis,* 54.

White, B. & Frederiksen, J. (1990). Causal model progression as a foundation for intelligent learning environments. In: W. J. Clancey & E. Soloway (Eds) Artificial intelligence and learning environments. Amsterdam, the Netherlands: Elsevier Science Publishers B. V.

Wilson, M. (2009, June*). The structured constructs model (SCM): A family of statistical models related to learning progressions*. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Wilson, M. & Scalise, K. (2006). Assessment to improve learning in higher education: The BEAR Assessment System. *Higher Education*, 52, pg 635-663.

Woolfolk, A. (2004). *Educational psychology, 9th edition*. Boston, MA: Pearson.