10-18-2016

Detecting Conceptual Change with Latent Transition Analysis

Glen Davenport
University of Connecticut - Storrs, glen.davenport@uconn.edu

Detecting Conceptual Change with Latent Transition Analysis

Glen Davenport, PhD

University of Connecticut, 2016

To uphold the core premise of cognitive diagnostic assessment, it is necessary to align different aspects of assessment design. The structure of the test, the measurement model, and the score reporting must align with each other and with the construct being targeted. Conceptual knowledge, as targeted by the Force and Motion Conceptual Evaluation (FMCE), can be modeled as a set of overlapping categorical states. As such, latent class analysis (LCA) is the most appropriate measurement model—where students have a probabilistic membership within classes that tend to match with observed mental models.

This dissertation focuses on one particular application of conceptual knowledge instruments: evaluating the effectiveness of instructional interventions within a controlled trials design. In these studies, students in different sections are taught using methods and assessed using the same instrument at pretest and posttest. Typically, the statistic of interest is the difference between the average changes in scores across the two time points. Given randomization and proper controls, researchers can use the results to make claims about which methods are more effective.

In the latent class framework, changes across time are captured with latent transition analysis (LTA) models, where transition parameters describe student posttest classes given membership in pretest classes. A multi-group LTA model can allow the transition parameters to vary across treatment groups while constraining measurement parameters. The difference in transitions from pretest to posttest across groups answers similar questions about the effectiveness of the instructional treatments, while providing more diagnostically relevant information.

The study described in this dissertation applies a multi-group LTA model to FMCE data from two large scale studies. The model was applied individually to each of the FMCE testlets, which focus on different concepts. This first application of latent class modeling to conceptual change was successful because many of the models converged, were fully identified, and provided interpretable results. Not every model converged, providing some clues about the limits of this method. The transition results agreed with conventional results that the instructional treatments were more effective than the more conventional instruction. However, it was difficult to find useful diagnostic information within the transition parameters.

Detecting Conceptual Change with Latent Transition Analysis

Glen Davenport

B.A., University of Maine, 2005

M.S., University of Maine, 2008

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

At the

University of Connecticut

2016

2016

APPROVAL PAGE

Doctor of Philosophy Dissertation

Detecting Conceptual Change with Latent Transition Analysis

Presented by Glen Davenport, B.A., M.S.

Major Advisor_____
H. Jane Rogers

Associate Advisor _____
D. Betsy McCoach

Associate Advisor _____
Hariharan Swaminathan

University of Connecticut

2016

Acknowledgements

There are many people who supported me through this project and to each I want to extend my sincerest gratitude. Dr. Jane Rogers and Dr. Betsy McCoach guided me through the dissertation writing process. Dr. Michael Wittmann provided the raw data used in this study and an enormous amount of professional advice. I want to thank Dr. Hariharan Swaminathan, Dr. Bianca Montrosse-Moorhead, and Dr. Christopher Rhoads, for reading my work and providing feedback.

My parents, Beth and Alan Davenport, gave me unending and infinite support. I want to thank Dr. Janel McDermott for her professional and personal support. Haley Autumn York was my biggest cheerleader. Katelyn Higgins was a constant force for good.

Thank you all so much.

Table of Contents

Table of Figures

Table of Tables

CHAPTER 1

INTRODUCTION

Concept inventories such as the Force and Motion Conceptual Evaluation (FMCE) are designed using rigorous cognitive research and are intended to diagnose student misconceptions. The design of the FMCE fits the philosophy of Cognitive Diagnostic Assessment (CDA), a subfield of psychometrics dedicated to providing specific information about student thinking. The authors of the FMCE used the results of misconceptions research, creating testlets that targeted specific misconceptions and including distractors that are attractive to students with particular states of conceptual knowledge (Thornton & Sokoloff, 1998). The connection between the cognitive research and the assessment design is very strong, giving the FMCE the potential to diagnose conceptual knowledge in detail. Unfortunately, practitioners typically use raw FMCE scores and do not have access to more sophisticated measurement models that would provide them with diagnostic information.

It is important, for score reporting and end-user purposes, that a measurement model matches the instrument's design and the cognitive model of the target construct. Since the FMCE targets conceptual knowledge, described in terms of categorical states, latent class analysis (LCA) appears to be the most appropriate measurement model for the FMCE. Latent class models classify individuals by their responses to assessment items, allowing for teachers and researchers to classify students by their thinking at the time of assessment. These models assume that each individual is a member of an unobservable subgroup within the population and that class membership determines the probability of particular responses. The fundamental notion of LCA, that of underlying categorical states that probabilistically determine item

responses, makes LCA the best candidate for accessing the diagnostic data that is washed out by raw scoring the FMCE.

Latent class analysis is advantageous in other ways as well. Individuals in LCA samples need not align perfectly with one class because the model accounts for off-class responses. If an individual responds in a way that implies membership in multiple classes, uncertainty in classification is represented by posterior probabilities of membership. Perfect classification means that each individual has a probability of one of belonging in their own class and zero probability of belonging in all the other classes. Latent class modeling accounts for the reality of non-discrete classification. Another advantage of LCA modeling is that the parameters can be used to create a scoring key that relates class membership and student responses. It is not necessary to estimate a separate LCA solution for each classroom to get estimates of class membership, as the formulae for posterior probabilities serve as a scoring key.

The biggest advantage of LCA scoring over continuous scoring is its ability to classify individuals using *combinations* of responses. Consider a four-item testlet where A is always the correct answer, B is the response chosen by students with the most common misconception, and C is another incorrect response. Conventional models dichotomize each item, awarding one point if the response is A and zero points if B or C. Some slightly more sophisticated measurement models attempt to account for distractors by dichotomizing them as well (Bao & Redish, 2006; Bradshaw & Templin, 2014). Under such models, a student giving AAAA would receive four *correct points* and zero *misconception points*, a student giving BBBB would receive zero and four respectively, and a student giving AABB would receive two correct and two misconception points. Latent class analysis, because it uses joint probabilities, has the power to

identify AABB and BBAA as distinct classes. This flexibility allows LCA to model population subgroups that are distinct but would give overlapping responses, such as BBCC and BBAA.

Davenport (2013) used latent class analysis on the testlets of the FMCE, treating multiple choice responses as categorical indicators, and found that the classes matched up with states of conceptual knowledge. Each testlet had one class of students that provided correct responses, one class of students that responded with the most common misconception, and classes of students that gave responses indicating dual or hybrid conceptions. A *hybrid conception* is one that mixes features of correct thinking and incorrect thinking but is itself a distinct way of thinking about the topic. A *dual conception* occurs when students learn the correct concept but have not yet eliminated the incorrect version of the concept. They may respond with the correct or incorrect version from one occasion to the next. These correct, incorrect, hybrid, and dual conceptions are *mental models*, alternate ways of thinking about the same topic, that are activated by item stems on the assessment.

Categorical scoring methods are a useful tool that should be added to the cognitive diagnostic assessment toolbox. They can join continuous models, multi-dimensional models, and diagnostic classification models, providing interpretive support for situations where student learning is best described as a set of categorical states. Davenport (2014) conducted an interview study with high school physics teachers, where the teachers were asked to interpret LCA-based score reports and provide feedback, and concluded that LCA scoring may be useful for providing diagnostic information for instructional purposes. While the case for instructional use is promising, it is unclear whether latent class measurement modeling is useful in research contexts. Specifically, it is unclear whether categorical scoring would be useful in the context of evaluating or comparing educational interventions, a common application of concept inventories.

Interventions are typically evaluated using controlled trial designs, where students are assigned to different instructional treatments and pretest-posttest assessments are used to compare the impact of the treatments. In science and mathematics education research, conceptual knowledge is often targeted as the most important outcome. As students learn science topics, they go through a process called conceptual change, shifting from incorrect conceptions to correct conceptions in a non-linear fashion. Conceptual change can be modeled by latent transition analysis (LTA), a longitudinal form of LCA that yields probabilities of transitioning from one latent class to another over time. The controlled trial, then, can be evaluated using a multi-group LTA model (mLTA), which would describe the conceptual change of students in each treatment group. Controlled trial studies often use linear regression models to compare performance across groups and time, but mLTA models have the potential to provide researchers with much more information about how student knowledge changed over the course of the study.

The purpose of this dissertation is to investigate the utility of mLTA in a controlled trial context. While the model has the potential to provide diagnostic information beyond regression analysis, it is an entirely untested method. The research literature includes a few examples of using LTA to describe conceptual change but none of the previous studies used multiple-choice responses as categorical indicator variables. Given the lack of prior applications, it was not clear (1) whether model estimation would converge on a solution (2) whether solutions would capture conceptual change in an interpretable way (3) or whether mLTA provides more information about students than linear regression. This dissertation is intended to be a proof-of-concept study, evaluating the potential of mLTA for use in controlled trial studies.

Using latent class models with data from two large scale physics interventions, this study aims to answer the following research questions:

1. What latent classes are present in the data set? How do the proportions vary across the two time points and the treatment conditions?

2. What does Latent Transition Analysis reveal about conceptual change in physics knowledge over a semester of instruction?

3. Can multi-group LTA detect significant differences between treatment groups?

4. Does multi-group LTA provide more information than a raw score comparison?

5. Do the answers to these questions vary across the FMCE testlets? Can differing results be attributed to the learning differences, testlet structure, or modeling issues?

The study is directly relevant to cognitive diagnostic assessment and psychometrics because it explores the properties of a statistical method that is rarely used as a measurement model, and never before in this specific form. If results of this study are promising, then LCA scoring and mLTA analysis can be tested in other contexts, applied to other conceptual knowledge tests, used to design novel assessments, and explored using simulation studies. If the line of research produces useful, defensible results, then categorical scoring could be applied widely in CDA applications where the underlying constructs are categorical.

The development of alternative measurement models may help to improve the validity as well as the utility of instruments by aligning more closely to the psychological structures being measured. LCA scoring also enables the use of modular assessments, where tests are composed of selections of relevant testlets. This allows instructors and researchers to measure student learning, and to compare those results against other studies, without being tied to the specific testlets on the FMCE. Finally, if the LTA and mLTA models function as expected, they could be used to explore the stability of conceptual knowledge using test-retest studies. This process

would give psychometricians a better sense of what is being measured by conceptual assessments.

This study will not evaluate the instructional treatments used with the original studies. Issues with fidelity, sampling, and data cleaning introduce a large amount of bias and uncertainty to the results. So, while it will not directly affect the field of physics education research (PER), this study will expand the current line of PER research on student assessment. It may also help to settle debates that have plagued concept inventories since the first was developed by PER researchers. Practitioners have debated how to analyze, interpret, and use assessment results. Some have argued that statistical analyses show that the assessments are invalid and provide no information at all (Heller & Huffman, 1995; Wallace & Bailey, 2010). I believe that this study can settle some of the debates by showing that the results are valid and interpretable if the correct measurement model is used. Sadler (1998) argued that, although these assessments are invaluable to content area research, raw scoring and even item response theory scoring of student responses are inappropriate. The line of research presented in this dissertation is consistent with the original intent of the FMCE authors, that "the FMCE was not originally designed to have results analyzed by a single number score." (Thornton et al., 2009; p. 2).

CHAPTER 2

LITERATURE REVIEW

This chapter describes previous research to provide background information and set the context of the current study. The first section describes two fields of study that provided a theoretical scaffold for the study: cognitive diagnostic assessment and modern schemes for valid test design. The next section gives background on the target construct, conceptual knowledge, and instruments that have been designed to measure it, concept inventories. After narrowing in on the history and format of the specific concept inventory used in this study, the chapter expands again to survey the literature for other examples of latent class modeling with conceptual knowledge. The research background of the latent class models appears in Chapter 3.

Theoretical Framework

Cognitive Diagnostic Assessment

Cognitive diagnostic assessment is an approach to test design that focuses on identifying and characterizing cognitive processes rather than ranking students on performance. Student behaviors such as selecting responses on multiple-choice items are taken as evidence for particular cognitive processes. This approach to assessment design was called for by prominent researchers for years (e.g., Messick, 1989), but was not a part of mainstream psychometric practice until the early 2000s. Anne Anastasi commented in 1967 that “those psychologists specializing in psychometrics have been devoting more and more of their efforts to refining techniques of test construction, while losing sight of the behavior they set out to measure” (Anastasi, 1967; p. 297).

Sternberg (1984) described the issue as a "sociological one" where institutional inertia kept test developers focused on ranking scales. I suspect that several factors fed into that inertia. First, diagnostic instruments must be based on the results of cognitive psychology research, a discipline that did not become popular until the 1980s. While the field now has a strong foundation overall, it takes years to explore each content area and describe how students learn within each discipline. Second, there were very few measurement models that could capture the information provided by diagnostic assessments. Finally, diagnostic assessments require a much more rigorous validation process, a greater investment of time and energy on the part of test developers. Leighton and Gierl (2007) described the validity issue as a "radical shift" in thinking and summarized by saying that "CDA requires us to pursue a rigorous program of validation, one that is focused on measuring the students' mental processes as they engage in test-taking behaviors and then using this information for improving students' opportunity to learn." (p. 7)

The turning point in cognitive diagnostic assessment appears to be the publication of *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001), a text commissioned by the National Research Council. The book summarized the progress that had been made to incorporate psychological research into test design. The authors called on researchers to continue their work on student learning and called on the testing industry to adopt CDA practices. In the years that followed, several developments allowed CDAs to enter mainstream psychometric practice. First, educational simulators and video games started providing researchers with enormous amounts of data that demanded more sophisticated measurement models. Second, studies showed that teachers wanted more diagnostic student data to help them provide targeted instruction (Huff & Goodman, 2007). Finally, perhaps most importantly, researchers developed a family of diagnostic classification models (DCMs).

DCMs use a series of categorical latent variables to represent small scale attributes that are part of a larger mental construct (Rupp, Templin, & Henson, 2010). For example, arithmetic problems could be decomposed into addition, subtraction, multiplication, and division processes. A student that can multiply consistently would 'have the multiplication attribute.' Instruments consist of items that require varying combinations of the attributes, so one item might require division and addition while another item might include multiplication and addition. The DCM takes student responses to the collection of items and calculates the probability that each student possesses each attribute. The attributes in the arithmetic example are pieces of factual and procedural knowledge, but the attributes could be heuristics (e.g., order of operations) or so-called non-cognitive traits (e.g., favorable attitudes towards mathematics).

Diagnostic classification models have been the subject of a large body of research in recent years, so much so that the term has become somewhat interchangeable with cognitive diagnostic assessment. A review by de la Torre and Minchen (2015) explicitly links the two, asserting that DCMs are necessary for CDA. While the importance and utility of DCMs is clear, I believe they should be considered one set of tools in a larger toolbox. Not all domains are best modeled by a set of dichotomous attributes—some constructs may be continuous or singly categorical. In this dissertation, a broader definition of cognitive diagnostic assessment is used: Research-based instruments designed to, in conjunction with the most appropriate measurement model, provide information about student thinking.

The CDA research community makes an assumption that CDAs lead to improved instruction and learning. While nearly all publications in the field assert that diagnostic data can be used to improve instruction, they do not cite research to show that the availability of diagnostic data has a beneficial effect. It is easy to imagine a hypothetical teacher who uses

diagnostic survey results to tailor instruction to a specific group of students, or a researcher who uses diagnostic survey results to make specific revisions to an intervention. However, without empirical results to support this assertion, the CDA community pays lip service to the impact of their own research. This is a significant omission at the center of CDA research that must be addressed in the coming years. A starting point might be the body of research on formative assessment (Black & Wiliam, 1998; Shute, 2008) or research on data within teacher preparation programs (e.g., Spitzer et al., 2010). These are fields of study that examine how teachers interpret and use student data.

Hoping to address this issue, I conducted an interview study with Advanced Placement physics teachers, asking them to (1) interpret score reports (2) explain how they would use the information provided by the reports and (3) provide feedback on the usefulness of categorical scoring for their practice (Davenport, 2014). My results confirmed the results of Huff and Goodman's (2007) teacher survey study. The participants approved of the diagnostic score reports and said they wanted more diagnostic data available to them. At the same time, it was unclear whether the teachers knew how to leverage that data effectively. Most were not able to say what changes they would make to instruction, given the FMCE scores of a hypothetical classroom. Unfortunately, I must also pay lip service to the impact of this study on student learning. Diagnostic assessments with latent class measurement models *could be* used by teachers to improve physics instruction.

Test Design and Validation Paradigms

Parallel to the development and growth of cognitive diagnostic assessments, psychometric experts have developed sophisticated approaches to test design and validation.

The conventional method, still used by most test developers, involves (1) defining the target domain, (2) collecting items from content experts, (3) pilot testing items to estimate item parameters, (4) removing items that have problematic parameters, show bias, or have differential item functioning, and (5) continuing until the items cover the breadth of the content domain. This procedure is effective for generating tests that accurately rank students in terms of their performance within a domain. However, this paradigm is not sufficient for generating more specific diagnostic data. Researchers and developers in CDA can turn instead to the design philosophies of Kane, Mislevy, and Wilson.

Kane's *argument-based* approach to validity (Kane, 1992; Kane, 2011) is influenced by Toulmin's (1958) model of inference. Toulmin said that when data is used to make a claim, the use of data must be supported by a warrant or defensible backing. Kane used this idea to give structure to the instrument validation process, where designers identify and test all of the pieces of the inferential chain of assessment. Kane's approach includes an *interpretive argument*: "a network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the assessment scores." (Kane, 2010; p. 8) It also includes a *validity argument* that defends each inference in the interpretive argument, essentially acting as the warrant of inference.

As an example, consider a placement test that divides incoming college students into those who are ready for calculus and those who are not prepared. The interpretive argument begins with unobservable psychological constructs and ends with how test scores will be used. It is a chain of reasoning that consists of dozens of claims, including the assertion that all test items are relevant to college mathematics, the assumption that there is only one value of the underlying construct that can produce each numerical score, and the inference that those students below the

cut score should not take calculus. The validity argument in this example would need to show, respectively, that the instrument has content validity, that the instrument is unidimensional and internally consistent, and that students below the cut score benefit from course work before taking college calculus. Such an instrument would have many more assumptions and inferences to defend, each requiring some amount of research. Building a complete validity argument requires a significant investment of time and resources, but is beneficial because it provides a complete picture of why the instrument should be used for its intended purpose.

Evidence-Centered Design (ECD) also focuses on the importance of assessment as an inferential argument, though it is more of a prescription for development than a process of validation (Mislevy, Steinberg, & Almond, 2003; Mislevy & Haertel, 2006). Assessment development with ECD begins with a *domain mapping* of the content area, cataloging the important skills and knowledge within the domain. The map includes the content itself as well as its hypothetical structure within the minds of students, allowing assessment designers to generate a *student model*. They then construct an *evidence model* that acts as a blueprint for the many inferences that will be made when test scores are interpreted. The evidence model explicitly connects items and tasks to skill and knowledge elements. It also details the statistical model that will be used to turn responses into numerical indicators. Finally, the *task model* is the set of items and activities that fit the specifications of the evidence model.

The main focus of ECD is the explicit link between test items, statistical models, and the model of student knowledge. It is a design paradigm that focuses on the fine-grained aspects of assessment. It details how inferences should be made about each piece of the test and the content domain, constructing a defensible argument for making inferences using full test scores from the ground up. The ECD process was developed by a team of researchers at the Educational Testing

Service, led by Robert Mislevy. It was used by assessment developers at the College Board for the redesign of the Advanced Placement exams (Bejar, 2010). The ECD process is prescriptive but allows for a wide range of measurement models, as long as those models match the model of student knowledge. Developing assessments with ECD is a labor-intensive process that requires collaboration with content experts at every stage but has the advantage of some amount of built-in validity. Each test item and each model parameter serves a specific pre-determined purpose and is tested individually during development.

The assessment system of the Berkeley Evaluation and Assessment Research (BEAR) program is a similar approach to design (Wilson & Sloane, 2000). Wilson's principles of design are expressed very clearly in Wilson (2008) along with the four most important components of the BEAR system. The first component is a *construct map* that explicitly lays out the performance expectations of students at different knowledge levels, with explicit links to learning and cognitive research. The second component is a set of items that link the elements of the construct map to instruction, pedagogy, and curricula. Next is the *outcome space*, which is how the assessment is embedded into the curriculum and how results are used to inform instruction. Finally, the *measurement model* takes results from the classroom and analyzes them in a way that allows for refinements of the construct map. Wilson diagrams these elements as a cycle where assessments are applied, analyzed, and reformed, always toward the purpose of making assessments representative of student knowledge and useful for teachers.

The BEAR system is most often associated with learning progressions, a model of conceptual knowledge, and a measurement model specifically for multiple choice items with ordinal response options. On such an assessment, each item has responses that correspond to specific, research-identified levels of student understanding. While it is most associated with

these specific assessments, the principles and values of the BEAR system are widely applicable. Wilson focuses on how the assessments will be used in the classroom and how collaboration with teachers facilitates the refinement of the assessments over time. He also discusses the explicit links that need to be made between student thinking, assessment construction, and assessment application.

The three approaches to test design, evaluation, and validation serve slightly different purposes and they prescribe different activities, but they share a common theme. Each approach emphasizes the connection between cognitive structures, test structures, and end-user application. These paradigms could be described as holistic approaches to assessment, focusing on parallelism across each level of inference. The current line of research uses a synthesis, greatly simplified, of the methods described above. The scheme in *Figure 2-1* illustrates a framework of cognitive diagnostic assessment that connects the important aspects of CDA in a linear fashion. It resembles Kane's argument-based validity in demanding that each step in the chain be evaluated. The scheme follows Mislevy's ECD in that it demands an explicit description of how each element aligns with each neighboring element. As with Wilson's BEAR system, the scheme emphasizes how the assessment will be used by teachers or researchers.

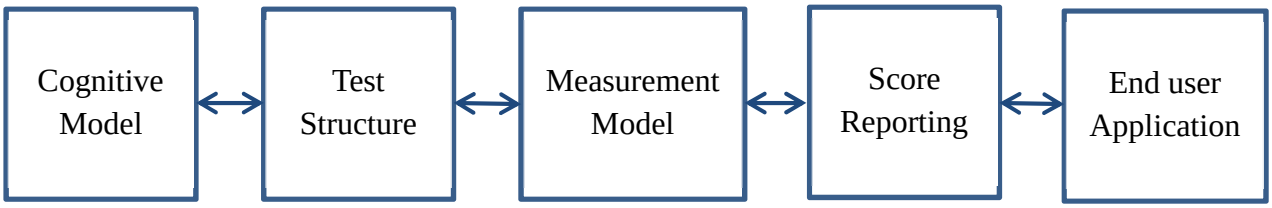


Figure 2-1. A simple scheme for developing effective and valid assessments.

This framework requires that (1) the test structure be capable of describing student thinking in a way that is consistent with cognitive research (2) the measurement model applied to the assessment must be consistent with the structure of the test (e.g. dimensionality and ordination) (3) the assessment score reports must be able to capture and then communicate the output of the measurement model so that (4) the end-user can make informed decisions based on valid inferences. If the elements align and are explicitly linked, then the decisions made by the end user are defensibly based on what is actually happening in the minds of the test takers.

The current study is an exploration of the Force and Motion Conceptual Evaluation within the context of this scheme. The instrument is based on rigorous cognitive research about student conceptual knowledge in introductory physics. The test is structured specifically to explore student conceptions, using testlets that target specific topics and response options that are attractive to students with specific misconceptions. The cognitive model and test structure are consistent, with links made explicit by the test developers and later researchers (Thornton & Sokoloff, 1998; Smith & Wittmann, 2008). However, all of the other elements in *Figure 2-1* are confused or absent for the FMCE. The instrument has two intended uses: helping teachers refine instruction and allowing researchers to compare instructional methods. The raw score measurement model that is most often used does very little to connect those functions with the cognitive model.

The framework in *Figure 2-1* must be applied separately to each context and purpose for the assessment. Davenport (2014) used latent class analysis as a measurement model for the FMCE, generating sample score reports and interviewing teachers about how they might use those reports. The goal of that study was to explore the inferential chain from ‘categorical conceptual knowledge’ at one end of the scheme to ‘refining instruction’ at the other. However,

the current study is aimed at the FMCE as a research tool for comparing instructional practices. This dissertation puts 'multi-group latent transition analysis' in the measurement model slot and evaluates the inferential chain that connects 'categorical conceptual knowledge' through to 'comparing instructional methods.' *Figure 2-2* names the elements in the chain of inference for the current study. The results may guide decisions about the next stages of research, perhaps looking at latent class scoring as a diagnostic assessment tool in a broader sense, in other contexts. Regardless of the results of this and follow up studies, I assert that it is essential that measurement models be selected for their alignment of the cognitive model and the end-user application. All too often, researchers select a model because it is sophisticated or flashy, rather than choosing the right model for the specific content and context.

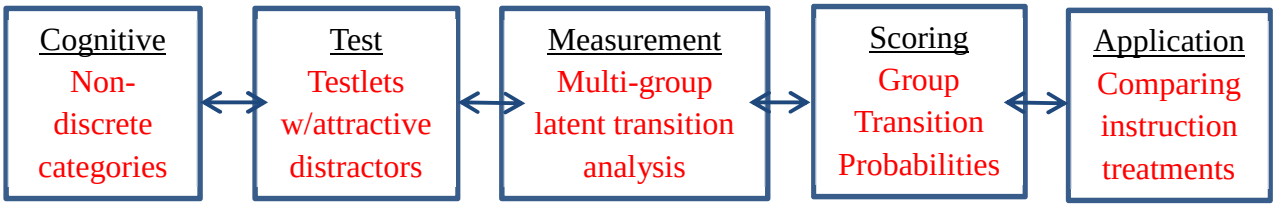


Figure 2-2. A diagnostic scheme specified for the FMCE as a comparison of conceptual change.

Conceptual Knowledge

Physics education research (PER) began, as a field, in the late 1970s and early 1980s, drawing heavily from constructivist thinking. Researchers found that students entered physics classrooms with beliefs, developed through day-to-day experience, that were incompatible with a Newtonian understanding of the world (e.g. Champagne et al., 1980). This led to a model of student learning based on *misconceptions* and *conceptual change* (e.g. Posner et al, 1982). One example of a misconception is the *impetus theory* of motion (Clement, 1982), that "the act of

setting an object in motion imparts to the object a force or 'impetus' that serves to maintain the motion." (McCloskey & Kohl, 1983; p. 147). Students, and a significant proportion of adults, believe that forces continue to affect motion even after the force is no longer acting on an object. People with this misconception believe that the impetus dies away as the object continues to move. A linear example of this is a hand tossing a coin into the air, where students often believe that there is still a force on the coin as it ascends, though the hand is no longer touching the coin. In a curvilinear example, an object is attached to a string and swung in circles. Students with the misconception believe that the object will continue to move in a curved path when the string is cut. They believe that the path of the object will gradually straighten as the effects of the circular motion die off.

Two decades of PER research identified and cataloged student misconceptions in physics, sometimes referred to as naïve concepts (McDermott, 1991; McDermott & Redish, 1999). Researchers found that students often learned to solve physics problems, allowing them to succeed in physics courses, without ever changing their misconceptions. They also found that, while it is possible to learn to solve problems and answer questions without a Newtonian understanding of motion, naïve concepts interfere with learning and make physics a much more difficult experience for most students. Unfortunately, misconceptions are difficult to change, likely because they are formed by real world experiences long before they enter the classroom. The constructivist philosophy of learning refutes the idea that students are 'blank slates' that acquire knowledge by simply listening to a teacher speak (e.g. von Glaserfeld, 1998). Constructivists believe that students already have ideas and it is the teacher's role to guide them as they build their own understanding of the world.

Hardy et al. (2006) looked at misconceptions surrounding objects floating or sinking in water and proposed a scheme for classifying naïve ideas. They said that the term *misconception* should refer to ideas that are always incorrect and easy to disprove. For example, the belief shared by many young students that small rocks float is always incorrect and can be demonstrated easily by dropping pebbles into a tank of water. However, there are other ideas that are true in most day-to-day contexts but are incorrect heuristics because they are not always true. Hardy et al. named these ideas *everyday conceptions* and provided as an example the idea that wood objects float and metal objects sink. The rule is very often true and not quickly disprovable by placing a metal object and a wooden object in a tank of water. The everyday conception can be addressed using specialized demonstrations, such as placing metal toy boats in the tank. Many of the ideas addressed by the FMCE are everyday conceptions. The impetus theory is an everyday conception because it accurately describes much of the motion that is observed in a world of friction and air resistance. For simplicity, this report refers to all incorrect ideas as misconceptions, though it is useful to keep in mind that the ideas are not entirely incorrect or useless. To varying degrees, the misconceptions measured in this study exist because they are useful in everyday life.

At its most basic, the misconceptions model of student learning is categorical, where students either have the correct conception or one of a set of misconceptions. Not surprisingly, the reality is more complicated. While it is possible for students to entirely abandon a naïve concept for the correct concept, many students change in ways that are not so discrete or linear. Demastes, Good, and Peebles (1996) explored conceptual change on the topic of evolution and natural selection and identified four patterns of conceptual change. A *wholesale* change occurs when a student simply abandons a misconception and internalizes the correct conception.

Incremental change occurs in stages, students internalizing one piece of the correct concept at a time. A *cascade* change occurs when the modification of a single idea allows for several other ideas to fall into place immediately. Finally, when a student learns the correct version of a concept but does not abandon the incorrect version, they form a *dual* misconception.

It can be very difficult to distinguish among the different forms of conceptual change described by Demastes, Good, and Peebles. These are processes that occur within the minds of students and are only observable through qualitative research methods. As an example, if a student has a misconception at one time point then responds correctly at the second time point, it is impossible to know whether the change occurred in a wholesale fashion or as a cascade. Indeed, it is impossible to know how whether wholesale changes occur as cascades unless the change actually occurs while under observation. Similarly, students with dual conceptions may give the correct answer after instruction, though they still have the incorrect idea in their mind. One common anecdote in physics education research illustrates dual conceptions by quoting a student who asked "So, on this survey thing, do you want me to answer what you want to hear or do you want me to answer what I really think?"

Vosniadou and Brewer (1992) studied first graders learning that the earth is a sphere and found that students can land on intermediary ideas as they move from misconceptions to correct ideas. Young children usually believe that the earth is a flat plane. When they are told that 'the Earth is round' they do not immediately picture a spherical world where life occurs on the outer surface. When asked verbally, students often repeat 'the Earth is round' to appease the grown-ups, often obscuring the fact that roundness may have been incorporated inaccurately into their model of the world. Some students imagine that the Earth is flat but circular, like a pancake. Others imagine a round world with a flat surface on top, as if a ball had been squashed flat on

one side. Some students visualize a 'fishbowl,' where a bottom hemisphere is stone and an upper hemisphere is clouds and sky. These students imagine the ground to be a flat surface across the middle of the sphere. Other students agree that the Earth is a round sphere, but do not know that they live on the Earth. They believe that they live on a flat, square surface and the Earth is one of the planets up in the sky. Vosniadou and Brewer show that by fifth grade, most students have a correct model, though some still cling to the hollow sphere—'fishbowl' –model of the world.

The term *hybrid conceptions* refers to ideas that share features of the correct concept and the common misconception but are distinct from either. Hybrid misconceptions are states of understanding that are internally consistent, making coherent—though not correct—sense. In contrast, dual conceptions may present coherently in one context but will not be consistent across time points or contexts. Additionally, there are other conceptual states where students get some pieces correct and some incorrect but not in a way that demonstrates a separate, coherent idea. Note that the terminology used here, such as hybrid conception, is not standardized across science education or psychological research.

The stages of conceptual change are described by the *learning progressions* framework in ways that are different from the conceptual change framework, though the two are not incompatible. Learning progressions describe the stages of accuracy and sophistication of student understanding regarding specific content (Duncan & Hmelo-Silver, 2009; Duschl, Maeng, & Sezen, 2011). These models of staged learning are often linked to grade level and the middle stages sometimes resemble hybrid misconceptions (e.g. Alonzo & Steedle, 2009). Students may take different paths through the progression as they learn (Wilson, 2008). Whether

mixed states of knowledge fit definitions of hybrid, dual, or mid-progression, they should be accounted for by the test structure and by the measurement model.

The fact that some conceptual changes are incremental and can generate hybrid conceptions implies that a concept is not a rigid, unified mental structure. Physics education researchers take the results of Vosniadou and Brewer as evidence for a knowledge-in-pieces model of student learning (diSessa, 1988). Also known as the *resources* model, this theory proposes that knowledge is stored as tiny pieces in the brain which are activated by stimuli and rapidly assembled to fit the context (Hammer, 2000). diSessa (1983) named some of the pieces *phenomenological primitives*, which are elements of knowledge that form during early childhood and are used to define simple heuristics. For example, the 'more is more' phenomenological primitive is a rule that works for many situations in day-to-day life: more force causes more motion, more fire is more hot, more cookies are more delicious. 'More is more' is a very fine-grained idea that is activated often and used in conjunction with other ideas to create conceptual understanding. There are many types of fine-grained knowledge, collectively known as resources, including pieces of declarative knowledge and epistemic resources (Louca et al., 2004). The resources model predicts that conceptual change can occur incrementally, that students can form unusual (hybrid) ideas by assembling pieces of knowledge incorrectly, and that conceptual knowledge is context specific.

In the 'Earth is round' example described above, the students have specific resources they access when answering the shape-of-earth question. They know that their experience of the world is flat and they know that the teacher said the world is round. When asked about the world, they activate both resources and assemble them in a way that makes sense, at least in the moment. The resources model describes the human mind as a sense-making machine that

rapidly assembles pieces, as needed, to satisfy the needs of the current context. A student missing the 'the place where I live is called the Earth' resource will take the 'my world is flat' and the 'the Earth is round' resources and assemble them in a way that reconciles all pieces and answers the question: "I live on this flat place and the Earth is a round ball up there." Other students know that their home world is called Earth and they find other ways to assemble their resources to answer the question.

Some researchers believe that the misconceptions model and the resources model are opposing, conflicting views of student thinking. I do not believe this is the case. The two models seem easily reconcilable. While it is clear that knowledge structures are not rigid or unified, neither are resources entirely independent. Resources associate with each other, particularly after being repeatedly co-activated. If the same resources are used often enough, they become highly associated and the knowledge structures become more rigid. This idea of resource *plasticity* (Sayre, Wittmann, & Donovan, 2007) shows how misconceptions and correct conceptions can appear to be rigid and unified while the process of learning reveals fragmented knowledge (Schneider and Hardy, 2013). While students are in the midst of conceptual change, they are learning new resources and trying to incorporate them into their existing knowledge. The result is that, when prompted, students can give inconsistent or unusual responses.

When activated and assembled, the resources form a *mental model* of the situation. The mental model, often coherent within itself, can be a correct, incorrect, or hybrid representation of a scenario. In this dissertation, the term misconception is used to mean an incorrect mental model that is commonly observed among students. The terms correct and hybrid refer to the mental models assembled when students answer FMCE items. My use of the term misconception is not a stance in favor of the misconceptions model or a stance against resources.

Whether conceptual knowledge is unified or fragmented, the result is the same at the time of assessment. Students form mental models for the duration of the time that they answer testlet items.

Concept Inventories

The first conceptual instrument to see widespread use was the Force Concept Inventory (FCI; Hestenes, Wells, & Swackhamer, 1992). The FCI followed from detailed research on student thinking that occurred in the 1980s and specifically targeted misconceptions surrounding the concept of 'force' (Halloun & Hesenes, 1985). Conventional physics assessments use, almost exclusively, items with algebraic problem solving. Students learn problem solving techniques, from lecture notes and homework problems, which allow them to answer items correctly without understanding the conceptual core of the content (e.g. Bagno & Eylon, 1997). The FCI used items that were entirely conceptual, divorcing the physics concepts from problem-solving skills. *Figure 2-3*. A sample item from the Force Concept Inventory. shows an example item from the FCI, relating to the curvilinear impetus misconception described in the previous section. If the string breaks at point P, the ball will follow path B as there is no longer any force to change its velocity. Students with the impetus misconception will select answer A because they believe that some motion has been imparted to the object, so it will continue moving as it had previously. They do believe that imparted motion dies off, so the path of the ball will begin to straighten out over time.

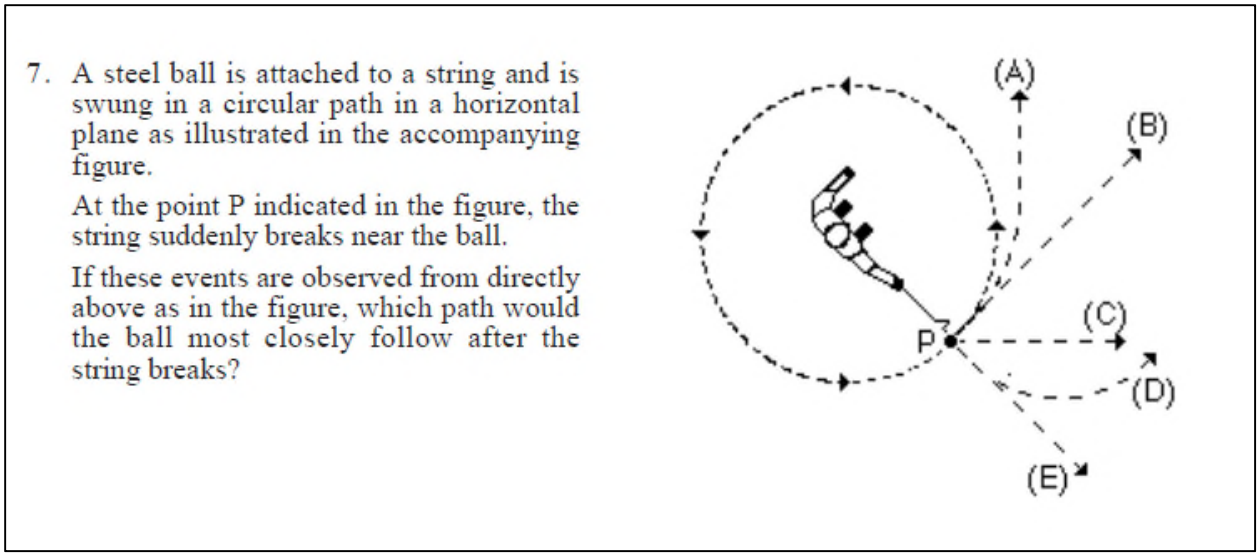


Figure 2-3. A sample item from the Force Concept Inventory.

The FCI set a standard for concept inventory construction by including *attractive distractors* in each item. These are responses that students with misconceptions are very likely to select. Answering FCI items requires "a forced choice between Newtonian concepts and common sense alternatives" (Hestenes, Wells, Swackhamer, 1992; p. 2). The phrase *common sense* is another term that some researchers use for misconceptions or naïve conceptions. The FCI has 30 items, the majority of which are grouped into testlets with common stems. The designers provided, in their 1992 publication, a mapping of item responses to incorrect student thinking. Their preliminary results, across 18 classrooms, showed that high school students scored near 50% at posttest. The authors pointed to the result as a warning sign that even honors and Advanced Placement students leave physics courses without a basic understanding of forces.

In the time since its publication, the FCI has become widely used by teachers and researchers to evaluate instruction. The instrument has been translated into more than 25 languages and has been used in many research contexts. FCI scores were found to correlate with proportional reasoning (Coletta & Phillips, 2005). The inventory was used to examine the

gender gap in physics (Dietz et al., 2012; McCullough, 2011). Overwhelmingly, though, the FCI is used to compare the effectiveness of pedagogical methods in changing student conceptual knowledge. A study of more than 6000 students, confirmed recently using data amassed from more than 50,000 students, shows that interactive-engagement methods are superior to traditional instruction by lecture (Hake, 1998; Von Korff et al., 2016). The Hake study was enormously influential, giving motivation to a growing movement away from of physics education reform.

The large scale studies confirm the FCI authors' finding that students in traditional classrooms score around 50% at posttest, reinforcing the point that something must change in physics education, but do not show which specific interactive-engagement methods are more effective than others. In the years that followed, researchers studied peer instruction (Crouch & Mazur, 2001), inquiry and problem based learning (Kirschner, Sweller, & Clark, 2006; Sahin, 2010), technology-infused studio physics (Cummings et al., 1999), and many others. The results of the research cannot determine a 'best' pedagogical method, though the results all tend to confirm that any interactive method is preferable to lecture-format instruction. This flurry of research, based on a cognitively rich assessment, was an influential enough that the FCI earned a place in the seminal text *Knowing What Students Know* (2001). Concept inventories were developed for other physics topics such as electricity (Maloney et al., 2001), energy (Ding, Chabay, & Sherwood, 2013), and wave mechanics (Tongchai et al., 2009). Concept inventories were developed in other disciplines, covering such topics as genetics (Smith, Wood, & Knight, 2008), evolution (Anderson, Fisher, & Norman, 2002), astronomy (Sadler et al., 2009), and statistics (Stone et al., 2003). In recent years, work on concept inventories has started to merge with work on learning progressions (Fulmer et al., 2014; Stains et al., 2001; Steedle & Shavelson, 2009).

From the beginning of concept inventories, researchers have debated their psychometric merit. In 1995, a series of papers debated the validity of the FCI on the grounds that an exploratory factor analysis did not find the expected factor structure (Huffman & Heller, 1995; Hestenes & Halloun, 1995; Heller & Huffman, 1995). Much of the debate circled around what factor analysis really means and how fragmented student knowledge can differ from expert knowledge. While the debate brought up interesting epistemological questions, it appears that there were methodological problems with the factor analysis itself. Hake (1998) recommended a calculation of gain scores that he called *normalized gain*, a statistic that became standard in PER though it provides no benefit over gain scores and is unnecessary in regression analysis. Other researchers have attempted to 'fix' normalized gain (Marx & Cummings, 2007) and warn against its use (Wallace & Bailey, 2010). Wang and Bao (2010) used item response theory to address the floor and ceiling effects that the FCI typically has. Sadler (1998) pointed out some specific failings of classical test and item response theories as applied to distractor-driven instruments and called for alternative measurement models. Henderson (2002) addressed common concerns that instructors and researchers had about the application of concept inventories, including the impact of grading the students on the FMCE and the testing effect of being exposed to the FMCE more than once. He found that neither effect impacted student performance. Bradshaw and Templin (2010) used the FCI as a pilot for their Scaling Individuals and Classifying Misconceptions model, a modified version of the diagnostic classification model that simultaneously estimates student IRT proficiency estimates.

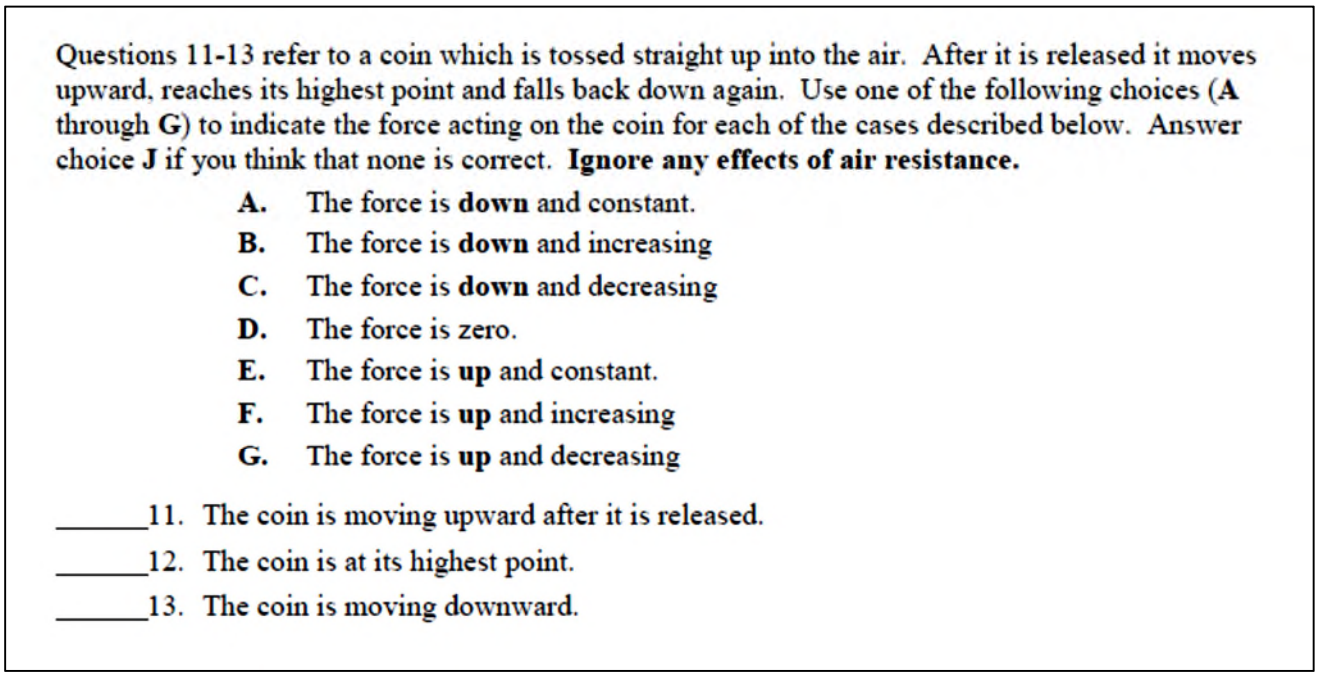
Wallace and Bailey (2010) titled their article *Do Concept Inventories Actually Measure Anything?* The title is indicative of a deep skepticism that permeates concept inventory research. It seems that some of the skepticism stems from unfamiliarity with statistical and psychometric

methods. The researchers conducting all of these studies were content area experts who dove into psychometric analysis with varying degrees of guidance from methodologist colleagues. It is not surprising that some analyses were performed incorrectly or that would be met with skepticism. Another cause for skepticism seems to be concern that student thinking cannot be represented appropriately with a numerical scale. The content experts connect with students regularly and are familiar with the categorical misconceptions that may or may not be made of fragmented pieces. Applying latent class modeling to concept inventories may relieve the skepticism surrounding their use by providing a statistical method that more directly represents student thinking.

The Force and Motion Conceptual Evaluation

The Force and Motion Conceptual Evaluation, presented in Appendix A, is a concept inventory that covers many of the same mechanics concepts as the FCI (Thornton & Sokoloff, 1998). The FMCE has a greater emphasis on graphical representations, asking students to interpret graphs of force, acceleration, and velocity versus time. This emphasis on graph interpretation is consistent with the authors' Real-Time Physics curriculum, which uses sensors and graphing software to help students engage with conceptually rich activities (Sokoloff, Laws, & Thornton, 2011). The testlet structure of the FMCE is much stronger than that of the FCI, where there are no independent items. Each testlet uses a scenario to target a particular misconception, such as the example in Figure 2-4, which asks about the net force on a coin as it is tossed into the air. One feature that sets the FMCE apart from other concept inventories is that the response options cover all possible responses. The eight possible responses shown in Figure 2-4 allow the students to fully express their mental models. The FMCE testlets straddle the line

between multiple choice and free response items. This design is one of the features that makes latent class analysis ideal for measurement modeling. Students are able to represent their full mental model using discrete values.



Questions 11-13 refer to a coin which is tossed straight up into the air. After it is released it moves upward, reaches its highest point and falls back down again. Use one of the following choices (**A** through **G**) to indicate the force acting on the coin for each of the cases described below. Answer choice **J** if you think that none is correct. **Ignore any effects of air resistance.**

A. The force is **down** and constant.
B. The force is **down** and increasing
C. The force is **down** and decreasing
D. The force is zero.
E. The force is **up** and constant.
F. The force is **up** and increasing
G. The force is **up** and decreasing

_____11. The coin is moving upward after it is released.
_____12. The coin is at its highest point.
_____13. The coin is moving downward.

Figure 2-4. A sample FMCE testlet.

Since its publication in 1998, the FMCE has been used to evaluate the effectiveness of the studio physics program (Cummings et al., 1999; Hoellwarth, Moelter, & Knight, 2005); evaluate the effectiveness of a massive open online course (Balint et al., 2015); examine gender bias (Kost, Pollock, & Finkelstein, 2009); compare specific pedagogical approaches for teaching Newton's Third Law (Smith & Wittmann, 2007); and compare the overall effect of interactive-engagement teaching against traditional lecture instruction (Von Korff, et al., 2016). Thornton et al. (2009) compared the FMCE to the FCI, both in terms of content and student performance. They found that the correlation of scores on the two instruments was strong, close to 0.8, though

the floor effect was more prominent for the FMCE. Smith and Wittmann (2008) examined the FMCE within a resources framework, matching student responses with likely mental models. The FMCE has not overtaken the FCI in terms of popularity, but it does see widespread use.

The FMCE has been the subject of a number of psychometric studies performed by content area experts hoping to bring more defensible analyses to their research. An exploratory factor analysis by Ramlo (2008) yielded inexplicable results, possibly explained by the choice of an orthogonal rotation method. Talbot (2013) addressed the problems inherent to gain scores using IRT and partial credit models as a solution. He found that the models were more appropriate and defensible, but that they did not add enough above and beyond raw scoring to justify their use. Huang and Mislevy (2010) used the Newton's Third Law testlet as an example of how to apply a measurement model, the Anderson/Rasch multivariate measurement model, to an existing instrument as a part of the Evidence-Centered Design process. Bao and Redish (2006) proposed an entirely new measurement model where responses are coded as 'correct,' 'misconception,' and 'other' and group performance is analyzed in terms of a two-dimensional vector. Each measurement model allows for some defensible inferences to be made, though latent class modeling seems to be the simplest way to capture the categorical mental model at the heart of each FMCE testlet.

The complete FMCE is presented in Appendix A, while Table 2-1 through Table 2-6 give summaries of six of the seven FMCE testlets. The seventh testlet, Velocity Graphs, is not part of the analysis of this study because the items are too easy for students, thus does not discriminate between high ability and low ability. The tables presented here also omit the items that are not typically scored, and which were not used in any of the latent class models in this study. In this section, I give a brief description of the observed response patterns and the conceptions they are likely to represent. The division of items into testlets is described in Chapter 4. The arrangement

used in this study is typical but not the only defensible arrangement. The division is supported by the results of an exploratory factor analysis reported in Appendix B.

Table 2-1 describes the Force Sled testlet of the FMCE, where students are asked to select the force that would cause the described motion of a sled across frictionless ice. The common misconception targeted by these items is that students simply conflate force with velocity. If the object is moving to the right and increasing in speed, they say the force must be to the right and increasing. Newtonian physics defines force as proportional to the *rate of change* of velocity. There are response patterns in results of this testlet that indicate hybrid conceptions where students internalize the directionality of force but still conflate the magnitude. These students know that the force must go in the opposite direction of motion if the object is slowing down but continue to say an increasing force causes an increasing velocity. Those that give the response pattern ABFGB assume a constant force is sufficient to slow down an object but that an increasing force is necessary to increase speed. They may believe there is friction on the ice that will 'help' with the slowing down, may know there is no friction but still believe it is easier to slow something down than to speed it up, or they may be reacting to the asymmetrical image of the sled (see Appendix A).

Table 2-1

Summary of the Force Sled Testlet

Testlet Stem:	These items ask students to select a verbal description of a force that matches the motion of a sled across frictionless ice in one dimension.
Item	
1	Moving to the right, speeding up
2	Moving to the right, constant
3	Moving to the right, slowing down
4	Moving to the left, constant
7	Moving to the left, slowing down
Response	
A	Force to the right and increasing
B	Force to the right and constant
C	Force to the right and decreasing
D	No force
E	Force to the left and decreasing
F	Force to the left and constant
G	Force to the left and increasing
J	No response is correct

The Reverse Direction testlet, summarized in Table 2-2, is a combination of three smaller testlets. Each testlet asks about the force on an object as it moves upwards, at the point it changes direction, and as it falls. The first set of three items asks about a toy car that has been pushed up a ramp, the second set about the force on a coin during a coin flip, and the final set asks about the acceleration of a coin during a coin flip. The common incorrect response again conflates velocity with force, though for many students it may be an expression of the impetus model of motion. This version of force and motion, described earlier, causes students to believe that the hand imparts motion to the coin and continues to affect the coin as it travels, though the effect wears off and gravity takes over. These students tend to select GDB GDB GDB on the

nine testlet items. In reality, there is only one force on the coin during a toss: the constant, downward force of gravity.

Table 2-2

Summary of the Reverse Direction Testlet

Testlet Stem:	These items ask students to describe the net force on an object as it moves upward, stops, and comes back down. The first stem describes a car pushed up a ramp, the second two stems describe a coin toss.
Item	
8	The **force** on a **car** as it moves up a ramp
9	The **force** on a **car** at the top of its motion
10	The **force** on a **car** as it moves back down the ramp
11	The **force** on a **coin** as it goes up in the air
12	The **force** on a **coin** at the top of its motion
13	The **force** on a **coin** as it falls back down
27	The **acceleration** of a **coin** as it goes up in the air
28	The **acceleration** of a **coin** at the top of its motion
29	The **acceleration** of a **coin** as it falls back down
Response	
A	The force is down and constant
B	The force is down and increasing
C	The force is down and decreasing
D	The force is zero
E	The force is up and constant
F	The force is up and increasing
G	The force is up and decreasing
J	None is correct

There are some interesting half-correct responses, where students know that gravity is the only force as the coin moves, but give answer D at the top of the trajectory (the response set is ADA ADA ADA). They assume that ‘it cannot be accelerating if it is not moving’ which may be an application of the ‘nothing is nothing’ phenomenological primitive. The idea that there must be motion for there to be acceleration can be entrenched and difficult to change. The ADA

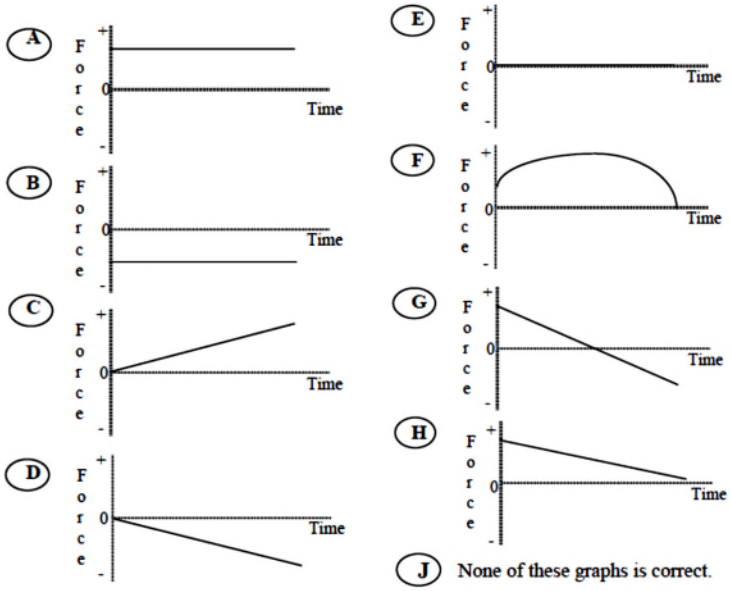
ADA ADA response pattern is a hybrid misconception where students combine the correct idea with an incorrect idea to form a third, distinct model of motion.

What is more interesting about the Reverse Direction testlet is the group of students who change their answers across the three mini-testlets. Some students respond GDB GDB AAA, indicating that, somewhere between Item 13 and Item 27, the correct mental model was primed and activated. Some of these students may genuinely see a difference between the force items and the acceleration items. In either case, the Reverse Direction testlet elicits responses that clearly represent a dual conception. The students have both concepts stored in their mind, or the pieces and connections that form both mental models, but they apply them inconsistently across time points or contexts.

The Force Graphs testlet, shown in Table 2-3, asks students to choose the graph of force vs. time that matches the described motion. The common misconception here is again a confusion of force and velocity, where students select an increasing force if the velocity is increasing. Students with Newtonian mental models select the graph that represents the rate of change of the velocity. These students tend to answer ACBDHF for the first six items and vary their responses the last item. These items would be correct if the item stem asked students to select velocity vs. time graphs. The correct response set is EAEBBGE.

Table 2-3

Summary of the Force Graphs testlet

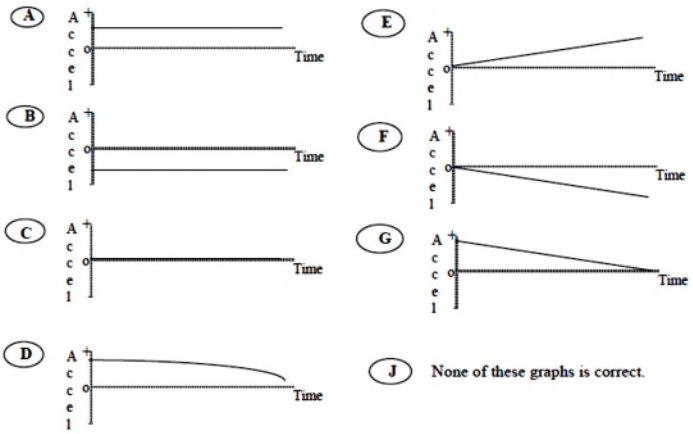
Testlet Stem:	These items ask students to consider a car moving in one dimension. Each item describes a motion and each response is a graph of force vs. time.
Item	
14	The car moves to the right, constant velocity
16	The car moves to the right, speeding up
17	The car moves to the left, constant velocity
18	The car moves to the right, slowing down
19	The car moves to the left, speeding up
20	The car moves to the right, speeds up then slows down
21	The car moves to the right, asks force after it is released
Responses	

The Acceleration Graphs testlet, abbreviated in Table 2-4, asks students to select the acceleration vs time graph that matches the described motion. The common misconception is to select the graph that matches velocity rather than acceleration, giving the response pattern EBGFA. Unlike the Force Graphs testlet, these items do receive some responses that imply a hybrid conception. In this case, the hybrid mental model has the students answering the constant

velocity items correctly and the other items as with the misconception. It is possible that these students learn the heuristic that 'constant velocity means no acceleration' without internalizing the full concept of one-dimensional motion. The corresponding response pattern is EGCFC. The correct responses set is ABCBC.

Table 2-4

Summary of the Acceleration Graphs testlet

Testlet Stem:	These items ask students to consider a car moving in one dimension. Each item describes a motion and each response is a graph of acceleration vs. time.
Item	
22	The car moves to the **right**, **speeding** up
23	The car moves to the **right**, **slowing** down
24	The car moves to the **left**, **constant** velocity
25	The car moves to the **left**, **speeding** up
26	The car moves to the **right**, **constant** velocity
Responses	

The Newton's Third Law testlet, abbreviated to Newton 3 or N3, is summarized in

Table 2-5. The first four items ask about a collision between a car and truck, while the last two items present a scenario where the truck has broken down and the car is giving it a helpful push. Students are asked to describe the forces between the car and the truck in each scenario. The correct answer is always that the two vehicles exert the same force on each other (EEEE AA). Students giving incorrect responses use one or both of two heuristics: The faster moving object exerts a bigger force or the bigger object exerts a bigger force (AFBB CB). Each is another misapplication of the 'more is more' phenomenological primitive. This testlet stimulates some students to reveal dual misconceptions. These students answer the first four items correctly and answer the last two items as if they have the common misconception. Clearly, they have both concepts stored in their minds but apply them inconsistently or differentially across contexts.

Table 2-5

Summary of the Newton's Third Law testlet

Testlet Stem:	The first four questions of the testlet use a stem where a car and truck have a head on collision. The last two questions ask about a situation in which the truck has broken down and the car is giving a helpful push.
Item	Asks students to compare the forces between a car and truck…
30	…if the two collide while moving the same speed
31	…if the two collide with the car moving much faster
32	…if the truck is standing still when the car collides with it
34	…if the truck is standing still and has the same mass as the car
36	…if the car is pushing the broken down truck, accelerating
38	…if the car is pushing the broken down truck, decelerating
Response	
A	The truck exerts a greater force
B	The car exerts a greater force
C	Neither exerts a force
D	The truck exerts a force but not the car
E	The truck and car exert the same forces
F	Not enough information to choose a response
J	No response is correct
	*Items 36 and 38 have the same responses in a different order

The Energy testlet, summarized in Table 2-6, presents a scenario where a child slides down a hill on a frictionless sled and reaches some speed at the bottom of the hill. Students are then asked to describe the speed or kinetic energy at the bottom of a steeper hill or a taller, but less steep, hill. The correct mental model is that only the height of the hill matters in a frictionless system because all of the potential energy will be converted into kinetic energy (BBAA). Some students, ignorant of the fact that energy and velocity are directly related, give different answers to 44 and 45 or 46 and 47. The more common misconception, built into student intuition through common experience, is that steepness influences speed (AACC). On a

frictionless hill, the final speed will be the same, but the acceleration will be greater. That is why steep hills 'feel' fast: humans tend to interpret acceleration as speed rather than velocity.

Table 2-6

Summary of the Energy testlet

Testlet Stem:	An image shows a child pulling a sled up to the top of a hill. The explains that after a frictionless slide down the hill, the sled has a speed v and kinetic energy E.
Item	Asks students to predict…
44	…the **speed** of the sled at the bottom of a **steeper** hill.
45	…the **kinetic energy** of the sled at the bottom of a **steeper** hill.
46	…the **speed** of the sled at the bottom of a **taller, less steep** hill.
47	…the **kinetic energy** of the sled at the bottom of a **taller, less steep** hill.
Response	
A	Greater than the original hill
B	The same as the original hill
C	Lesser than the original hill
D	Not enough information
J	None are correct

Latent Class Modeling and Conceptual Knowledge

There have been several applications of latent class modeling to diagnostic measurement. A very early example from psychiatry used latent class analysis to identify schizophrenic patients based on responses (Young, 1982). An early application of LCA to educational testing was to use the sequence of responses provided by students on a computerized test where students were instructed to 'keep answering until you get it right' (Wilcox & Wilcox, 1988). Kim (2005) used a latent Markov model to examine changes in student knowledge over time, similar to the current study except that the model only used one item at each time point. Schneider and Hardy (2013) coded responses on a 'floating and sinking' conceptual survey to misconceptions, everyday conceptions, and scientific conceptions. They used tallies of responses in each

category to perform a latent profile analysis (LPA), which is LCA with numerical scale indicators instead of categorical responses. Another study of student understanding of rational numbers used raw scores on three subscales to perform LPA and latent transition analysis (McMullen, 2015). One research team specifically looked for conceptual change over time, but their method used IRT item parameter estimates as indicator variables in latent transition analysis (Cho et al., 2010, 2011, 2013).

There have been some attempts to use latent class modeling in conceptual physics contexts, though the strategies used seem to vary widely from those in the current study. Steedle and Shavelson (2009) used latent class analysis to try to identify student misconceptions of force and motion. Unlike the current study, they coded student responses by learning progression level and used those levels as indicator variables in the measurement model. Fulmer et al. (2014) used latent class analysis with FCI data and found four classes with different mean scores. The authors used 20 items in the model, rather than targeting a specific topic, and did not specify whether the items were entered as categorical responses or as dichotomous values. The authors also did not try to characterize the classes in any way other than to compare scores. There are no published studies that use multiple choice responses from item testlets as indicator variables in LCA models to examine conceptual change.

Preliminary studies with FMCE and LCA used a sample of 1800 matched, complete responses to the FMCE at pretest and at posttest (Davenport, 2013). A set of latent class analyses, one for each of the testlets, used the responses themselves as input nominal variables. Analyses were performed in LatentGold Version 4.5. The optimal number of latent classes for each analysis was selected using the Bayesian information criterion (BIC). The classes that emerged were largely interpretable, always including one class of Newtonian thinkers and one

class of students with the common misconception. Some testlets produced classes that could be interpreted as hybrid or dual conceptions.

Davenport (2013) used the results of the LCAs to design classroom-level score reports. Table 2-7 is an excerpt from a pretest score report, showing that each student received a color coded label for each of the testlets. The labels were assigned according to the greatest posterior probability of membership, described in Chapter 3. Green indicated the correct answer class, red represented the common misconception class, while yellow indicated hybrid and dual conception classes, and white was assigned to 'other' classes. The goal of these reports was to provide teachers with diagnostic information that they could use to inform instruction.

Table 2-7

Excerpt from a sample LCA score report

Name	Raw Score	Force Sled	Reverse Direction	Force Graphs	Acc. Graphs	Newton Three	Velocity Graphs	Energy
Student 1	18	FS3	RD2	FG1	AG5	N1	VG1	E1
Student 2	5	FS1	RD1	FG1	AG3	N1	VG2	E1
Student 3	26	FS1	RD4	FG1	AG1	N3	VG1	E2
Student 4	19	FS1	RD5	FG2	AG1	N3	VG1	E4
Student 5	8	FS1	RD3	FG1	AG2	N4	VG1	E1
…								

Davenport (2014) conducted an interview study with 17 Advanced Placement physics instructors to determine whether teachers could interpret and apply the data in the reports. The majority of the teachers said that the LCA reports were easier to interpret than analogous raw score reports and a majority said that they wanted access to more diagnostic information. However, of the 17 teachers, only two could actually say how they would use the data. So, while there is a demand for diagnostic data in education the benefit is unclear. Meanwhile, the

question remains whether latent class scoring is useful in research contexts such as comparing curricula in large-scale controlled trial studies.

CHAPTER 3

LATENT CLASS MODELING OVERVIEW

This chapter presents the research background and general mathematical description of the latent class models used in the current study. The chapter begins with a brief introduction to latent class analysis and then dives into its history and mathematical specification. The next sections describe how the models are estimated, how practitioners select the preferred number of latent classes, and introduce a number of issues that are unique to latent class modeling. Finally, the chapter describes the longitudinal extension of LCA, latent transition analysis, and the multi-group extension of latent class modeling.

Introduction to LCA

Latent class analysis (LCA) is an exploratory measurement procedure that is used to identify subgroups within a larger population. It is analogous to exploratory factor analysis (EFA) in assuming that latent, or unobserved, variables influence the values of a set of observed variables. The difference between the two exploratory methods is that LCA posits a single, categorical latent variable rather than a set of normally distributed, continuous, latent variables. Figure 2-1 shows the path diagram of a simple LCA with circular elements for latent variables, square elements for observed variables, and an arrow for each set of parameters. The observed variables, often referred to as *indicator variables*, are assumed to be conditionally independent. They are independent from one another except through the common influence of the latent class variable. These models assume a causal relationship from class to response, though the model estimation process uses the observed responses to identify classes.

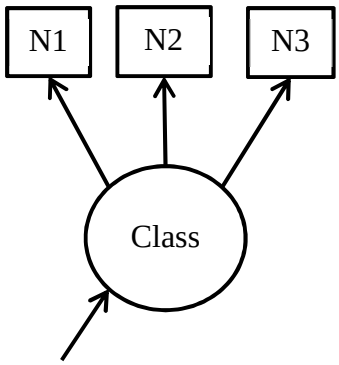


Figure 3-1. Path diagram of a latent class model.

Latent class modeling is appropriate in contexts where the underlying construct is some categorical state, a qualitative difference among individuals rather than a quantitative difference (Ruscio & Ruscio, 2008). Moreover, the underlying difference must impact the responses that individuals give to an assessment, though not in a way that is obvious to inspection. As an example, consider Table 3-1, which presents the ten most common response patterns on the Energy testlet of the FMCE. It is likely that student conceptions, assumed to be categorical states, drive student responses. However, it is very difficult to identify the conceptual states by looking at the response patterns. The table contains too much information to interpret class structure by eye, especially considering the dozens of response of response patterns not included in the table.

Table 3-1

The ten most common response patterns to the FMCE Energy testlet by students at pretest

Pretest Energy Testlet	
Response Pattern	Proportion of students
BBAA	0.13
AADD	0.14
AACC	0.11
ABCB	0.04
AAAA	0.03
BBBB	0.02
AABB	0.02
ABCA	0.02
ABDA	0.02
ABDB	0.02
…	…

A LCA produces a set of *item response probabilities* that give the probability of each response on each item, given membership in each class. The item response probabilities, which are estimated parameters in the model, can be used to describe and make inferences about the latent classes. Table 3-2 gives probabilities from a LCA of pretest responses to the Energy testlet, providing an example of this interpretation of parameters. In this study, classes are described in terms of their most likely responses, which appear at the top of each column in Table 3-2.

Table 3-2

Item response parameters, as probabilities, to a six class LCA solution of the pretest data from the Energy testlet of the FMCE.

	Latent Class					
	1 AADD	2 aa**	3 aaBB	4 AB**	5 BBAA	6 AABB
Item 44						
A	**0.93**	**0.62**	**0.64**	**0.85**	0.01	**1.00**
B	0.00	0.20	0.34	0.08	**0.99**	0.00
C	0.01	0.11	0.01	0.03	0.00	0.00
D	0.06	0.05	0.01	0.03	0.00	0.00
Item 45						
A	**0.89**	**0.59**	**0.57**	0.16	0.00	**0.97**
B	0.03	0.11	0.40	**0.80**	**1.00**	0.02
C	0.01	0.19	0.04	0.03	0.00	0.00
D	0.07	0.09	0.00	0.01	0.00	0.00
Item 46						
A	0.00	0.45	0.00	0.18	**0.97**	0.00
B	0.00	0.19	**0.97**	0.02	0.01	0.02
C	0.02	0.24	0.03	0.42	0.01	**0.98**
D	**0.97**	0.08	0.00	0.35	0.00	0.00
Item 47						
A	0.04	0.44	0.05	0.37	**0.95**	0.02
B	0.03	0.12	**0.81**	0.46	0.05	0.02
C	0.10	0.27	0.14	0.09	0.01	**0.95**
D	**0.83**	0.12	0.00	0.08	0.00	0.00

This dissertation uses a labeling scheme with capital and lower case letters to give a sense of the strength of response probabilities. Capital letters indicate that the modal probability is greater than 0.7, lower case letters indicate that the probability of the response is between 0.5 and 0.7, while asterisks indicate that no response had more than a 50% chance of being selected by members of that class.

Modal response patterns allow for interpretation of what the students may have been thinking when they selected their responses. For example, the BBAA group in column 5 is likely to be a group of students who have a correct understanding of potential energy because they are highly likely to select the correct answers. The labeling scheme also highlights the probabilistic nature of LCA models. As with factor analytic models, LCA assumes that there is no error in the latent variables, though there is 'error' at the individual level. Students in a particular class are not assumed to be completely consistent in their responses with other students in the same class.

History and Specification of the LCA model

Latent class analysis is attributed to Lazarsfeld and Henry (1968) who generated the mathematical basis for LCA before methods were available to estimate the parameters of the model. Goodman (1974) developed an estimation method to find the parameters that maximize the likelihood of the model. His method was a special case of the expectation maximization (EM) algorithm that is now commonly used to estimate LCA models (Dempster, Laird, & Rubin, 1977). The latent class models initially used probability parameters but were shown to work in a log-linear framework (Haberman, 1979, Formann, 1982; Hagenaars, 1998), which is the parameterization used in the current study. Dayton and Macready (1988) added covariates to latent class models, allowing for exploration of class membership in terms of other observed variables. The simple latent class analysis was extended to a longitudinal form known as latent transition analysis (LTA; Bye & Shechter, 1986; Collins & Wugalter, 1992).

A mathematical description of LCA begins with the contingency table, the complete set of all possible response patterns. A contingency table has W cells, where W is the product of the number of items and the number of responses for each item. The Energy testlet has four items, each with five possible responses, so $W_{Energy} = 5^4 = 625$ cells. The table is populated by N

individuals, each increasing the frequency by one in the cell that matches their response pattern. A contingency table is referred to as sparse if the number of individuals is small compared to the number of possible response sets. A ratio of W/N = 5 is desirable for the purposes of model identification and estimation (Read & Cressie, 1988).

When all latent and observed variables are nominal, as in the current study, LCA models reduce to a set of simultaneous multinomial logistic regressions. Without continuous variables, the model does not require any slope terms and so produces no variance-covariance matrix. The estimated model includes only intercept terms. Note that, because of this difference between LCA and typical structural equation models (SEMs), the arrows on path diagrams refer to mean structure parameters rather than regression weights and error variances.

A latent class analysis model relates a categorical latent variable X, which takes specific values c that range from 1 to *C*, to a set of observed dependent variables, denoted Y_1 through Y_I. When observed variables are nominal, each dependent variable Y_i can take specific values $\underline{r}$, from 1 to *R*. A LCA model consists of prevalence parameters and measurement parameters, represented by arrows in Figure 3-2. Note that each arrow in the diagram represents several parameters. *Prevalence parameters*, written as γ_c, determine the probability of being a member of each latent class and appear in path diagrams pointed towards the latent variable. The *measurement parameters*, also called item response parameters, are denoted by $\rho_{Yi,r,c}$ and represented graphically as arrows pointing from the latent variable to the observed variables.

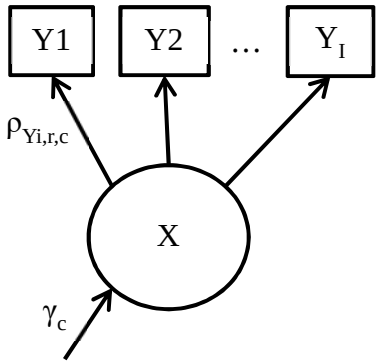


Figure 3-2. Path diagram for a latent class analysis model.

Each $\rho_{Yi,\,r,\,c}$ determines the probability that members of class *c* will give response r to item Y_i. The model assumes that all individuals are members of one of the classes, so the γ values sum to one across classes, and that all individuals select one response, so the response probabilities sum to one for a given item and class.

$$\sum_{c=1}^{C} \gamma_c = 1. \qquad \sum_{r=1}^{R} \rho_{r|Yc} = 1. \tag{3-1 and 3-2}$$

Given the probability that members of a class will choose a particular response, $P(Y_i = r \mid X = c) = \rho_{Yi,r,c}$, it is possible to calculate the probability of observing a particular response pattern, where a set of variables **Y** take on a set of observed values ***r***. The probability is given as the product of each applicable item response probability:

$$P(\boldsymbol{Y} = \boldsymbol{r} \mid X = c) = \prod_{i=1}^{I} \rho_{Yi,r=\mathbf{r}i,c} \tag{3-3}$$

where $\boldsymbol{r}_i$ is the value of element *i* of the ***r*** vector. To illustrate how equation 3-3 functions, consider class 3 shown in Table 3-2. Students in class 3 are most likely to give the response pattern AABB on the Energy testlet, but also have some probability of giving BBBB. That probability can be calculated using equation 3-3, by multiplying the item response probabilities that match the specified vector of responses. The probability of a class 3 student responding A to item 44 is 0.64, responding A to item 45 is 0.57, responding B to 46 is 0.97, and responding B to item 47 is 0.81. The product of these values is (0.64)(0.57)(0.97)(0.81) = 0.29. The analogous product for the BBBB response pattern is (0.34)(0.40)(0.97)(0.81) = 0.11. Students in the class have a 0.29 probability of giving AABB and a 0.11 probability of giving BBBB.

Incorporating the prevalence parameters allows for the calculation of the overall probability of observing a response pattern. First, it is necessary to calculate the joint probability of being in a class *c* and giving response pattern ***r***. Joint probabilities are given by P(AB) = P(A)P(B|A). In this case the probability of being in class *c* is γ_c and the probability of a response pattern given membership in class *c* is equation 3-4. So the joint probability is calculated as:

$$P(\boldsymbol{Y} = \boldsymbol{r}, X = c) = \gamma_c \prod_{i=1}^{I} \rho_{\mathrm{Yi,r=ri,c}} \tag{3-4}$$

The total probability of observing response pattern ***r*** is the sum of each of the joint probabilities across all of the classes. One could also say that it is the sum of each class's probability of giving **r**, weighted by the prevalence of each class. The weighted sum, shown in equation 3-5, predicts the distribution of frequencies across the contingency table.

$$P(\boldsymbol{Y} = \boldsymbol{r}) = \sum_{c=1}^{C} \gamma_c \prod_{i=1}^{I} \rho_{\mathrm{Yi,r=ri,c}} \tag{3-5}$$

The likelihood function for a set of data is the product of the probabilities of all observed response patterns given the parameters of the model, i.e.,

$$L(\boldsymbol{Y}|\boldsymbol{\theta}) = \prod_{j=1}^{N} P(\boldsymbol{Y}_j|\boldsymbol{\theta}) \tag{3-6}$$

where $\boldsymbol{\theta}$ is the full set of model parameters and $\boldsymbol{Y}_j$ is the response pattern of individual *j*.

Equations 3-1 through 3-6 describe a latent class analysis performed using a probability parameterization, where each parameter directly represents the probability of a response. Many software packages, including Mplus, use a loglinear parameterization (Haberman, 1979). Each loglinear parameter α is the log-odds of a particular response versus the reference response.

$$log\frac{P(Y_i = r|X = c)}{P(Y_i = R|X = c)} = \alpha_{Yi,r,c} \tag{3-7}$$

Mplus automatically assigns the last category to be the reference level of a categorical variable. The reference response does not have a parameter itself, so loglinear LCA models have fewer overall parameters than LCA models with a probability parameterization. There are $CI(R-1)$ measurement parameters and $C-1$ prevalence parameters in an ordinary LCA model, where *C* is the number of classes, *I* is the number of items, and *R* is the number of responses for each item.

The α parameters can be converted into item response probabilities, equivalent to ρ parameters, by calculating the probability that members of class *c* will give the reference response *R* of item Y_i, shown in equation 3-8.

$$P(Y_i = R|X = c) = \frac{1}{1 + \sum_{r=1}^{R-1} e^{\alpha_{Yi,r,c}}} \quad (3\text{-}8)$$

Multiplying the probability of the reference response by $e^{\alpha_{Yi,r,c}}$, the exponentiated form of a measurement parameter, gives the probability of observing response *r* on item Y_i. Table 3-3 illustrates the conversion from loglinear parameters to item response probabilities using a single class from the Energy testlet example.

Table 3-3

Calculation of item response probabilities from loglinear measurement parameters for the aaBB class (Class 3 in Table 3-2) of the Energy testlet, where the reference response is option J

Item	Response	Parameter α (log odds)	e^{α}	Calculation of P(J)	Probability $P(J)\cdot e^{\alpha}$
PRE44	A	5.64	280.33		**0.638**
PRE44	B	5.00	148.41	$\mathbf{\Sigma e^{\alpha}} = 438.1$	0.338
PRE44	C	1.81	6.12	$\mathbf{1+\Sigma e^{\alpha}} = 439.1$	0.014
PRE44	D	1.19	3.28	$\mathbf{(1+\Sigma e^{\alpha})^{-1}} = .0023$	0.007
PRE45	A	15.00	3.3E5		**0.565**
PRE45	B	14.64	2.2E6	$\mathbf{\Sigma e^{\alpha}} = 5.8E6$	0.396
PRE45	C	12.29	2.2E5	$\mathbf{1+\Sigma e^{\alpha}} = 5.8E6$	0.038
PRE45	D	8.99	8054	$\mathbf{(1+\Sigma e^{\alpha})^{-1}} = 1.7E-7$	0.001
PRE46	A	-15.00	3.06E-07		0.000
PRE46	B	15.00	3.3E6	$\mathbf{\Sigma e^{\alpha}} = 3.4E6$	**0.970**
PRE46	C	11.51	99807	$\mathbf{1+\Sigma e^{\alpha}} = 3.4E6$	0.030
PRE46	D	-15.00	3.1E-07	$\mathbf{(1+\Sigma e^{\alpha})^{-1}} = 3.0E-7$	0.000
PRE47	A	2.70	14.80		0.048
PRE47	B	5.52	248.39	$\mathbf{\Sigma e^{\alpha}} = 308.2$	**0.806**
PRE47	C	3.79	44.07	$\mathbf{1+\Sigma e^{\alpha}} = 309.2$	0.143
PRE47	D	-15.00	3.1E-07	$\mathbf{(1+\Sigma e^{\alpha})^{-1}} = .0032$	0.000

Once the item response probabilities are estimated, Bayes' theorem can be used to predict class memberships given a particular response pattern. The theorem, written algebraically as P(A|B) = P(B|A)P(A)/P(B), is adapted to LCA contexts as equation 3-9. The quantity $P(\mathbf{Y} = \boldsymbol{r}|X = c)$ is given in equation 3-8, the quantity $P(\mathbf{Y} = \boldsymbol{r})$ is given in equation 3-5, and $P(X = c)$ is a prevalence parameter. Substituting those values into 9 gives the probability that an individual with response pattern **r** is a member of class *c*. This value is known as a *posterior probability*

and gives researchers the power to predict unobservable attributes of individuals using estimated model parameters.

$$P(X = c|\boldsymbol{Y} = \boldsymbol{r}) = \frac{P(\boldsymbol{Y} = \boldsymbol{r}|X = c)P(X = c)}{P(\boldsymbol{Y} = \boldsymbol{r})} \quad (3\text{-}9)$$

Model Estimation, Convergence, and Identification

All latent class models in this study were estimated using an expectation maximization method (EM). The EM algorithm begins with random starting values for all of the γ and α parameters and iterates through two phases of estimation to converge on a set of best parameter values. The first phase, the expectation step, applies the parameter starting values to the observed data to obtain expected values of latent variables, and then uses those values to generate a weighted 'training set' of the indicator variables (Agresti, 2012). The maximization phase uses the expected latent class values and the training set of observed variables to generate new parameter values. Each iteration generates parameter values that increase the value of the likelihood function of the model. Most software packages use EM algorithms that maximize the log of the likelihood (LL), which shares maxima at the same locations as the likelihood, and continue iterating until the change in LL across iterations reaches some minimum criterion. This process is known as converging on a *solution*, a complete set of parameter estimates.

One complication of estimating latent class models is that the likelihood function may have local maxima. It is possible for the EM algorithm to converge on a solution that is not the best fitting solution. Figure 3-3 shows possible shapes of a likelihood function for a model with one parameter θ, some of which have local maxima. The two plots in the bottom row present

unique challenges, the first being many solutions that appear almost equally likely and the second being a model with no single maximum likelihood solution. Running many replications with different, random starting values mitigates the risk of converging on a locally maximal solution. Software packages such as Mplus allow the user to specify a number of starts and finishes for model estimation. For each start, the software picks random starting values for each parameter and runs through ten EM iterations. The most promising replications, those with the greatest likelihood after ten iterations, are selected to run to convergence.

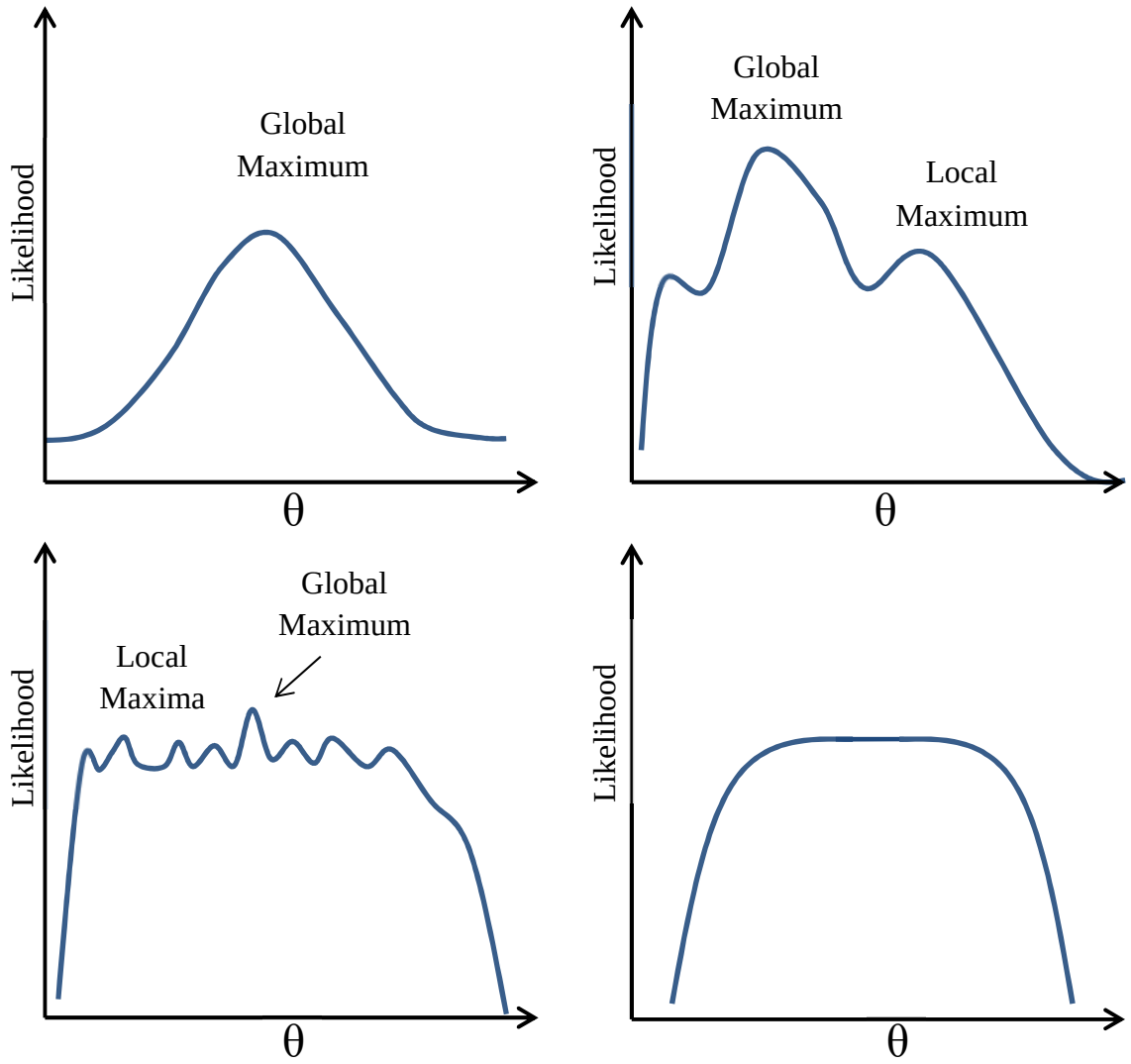


Figure 3-3. Possible geographies of the likelihood function.

Table 3-4 is an example of the convergence summaries generated by Mplus for every latent class model. It displays the LL values, offering some sense of whether the global maximum was reached, and seed numbers that can be used to explore the individual solutions. In this example, the solution with the greatest log likelihood was generated by ten of the twenty converged solutions. This is a strong indication that the best solution is the one with the -15705 log likelihood. Collins and Lanza (2010) recommend, as a rule of thumb, to be skeptical of a solution if its likelihood was not reached by more than one replication.

Table 3-4

Loglikelihood values, random number seeds, and replication number of the converged solutions of a six-class model of the Energy testlet.

LL	Seed	Rep. No.
-15705.6	341041	34
-15705.6	804561	59
-15705.6	551639	55
-15705.6	364676	27
-15705.6	432148	30
-15705.6	415931	10
-15705.6	963053	43
-15705.6	475420	71
-15705.6	27071	15
-15705.6	966014	37
-15705.7	268217	83
-15749.1	131856	90
-15749.1	573096	20
-15749.1	544048	87
-15749.1	372176	23
-15749.3	902278	21
-15760.6	957392	79
-15762.6	576596	99
-15764.0	136842	58
-15770.1	754100	56

Another important issue of latent class modeling is that of model identification, which means, in general terms, that there must be enough information available in the data to determine parameter estimates. LCA models can have identification problems that rise from the structure of the data itself, particularly when the number of individuals in a class is very small. There must be enough information within each class to estimate the parameters for that class. Model identification is impacted negatively by sparseness, overlapping classes, and a large number of latent classes. Identification issues can be addressed by collecting a larger sample, reducing the number of parameters by removing extraneous items, or constraining parameters so that they do not need to be estimated. Berzofsky et al (2014) showed that identification issues can also be the result of local dependence among items. They used simulations to show that items with relationships beyond that described by the latent class variable tend to cause local solutions, and cause weak identifiability in cases where the contingency table is sparse.

Model Fit and Selecting the Number of Classes

The fit of a latent class model is assessed using the likelihood and its associated information criteria. Unfortunately, there are no accurate tests of absolute model fit, particularly when the contingency table is sparse. However, it is possible to use relative measures of fit to select the most appropriate model. If two models are nested, the likelihood ratio test can determine whether the less constrained version of the model fits statistically significantly better than the more constrained version. This test is performed by subtracting the log likelihood values of two nested values and multiplying the difference by two. The resulting quantity is distributed as a chi-square with degrees of freedom equal to the difference in the number of estimated parameters.

Selecting the most appropriate number of classes is a crucial part of latent class modeling and one that involves many decisions on the part of the researcher (Collins & Lanza, 2010). Unfortunately, the likelihood ratio test cannot be used to compare LCAs with different numbers of latent classes because the models are not nested. In this case, the best statistical tool for selecting a number of classes is the Bayesian information criterion (BIC; Nylund et al, 2007). The formula for calculating this value is $BIC = -2LL + \log(N) \cdot P$, where LL is the log of the likelihood, N is the sample size, and P is the total number of estimated parameters. Lower BIC values indicate better fitting models. Researchers sometimes use the rule-of-thumb that a ten point difference in BIC indicates a significantly better fitting model (Kass & Raftery, 1995). In general, latent class models with sequential numbers of classes will have decreasing BIC values until a minimum is reached and the value begins to climb. The estimated model with the minimum BIC is likely to be the best solution, though researchers should also consider convergence, identification, and the properties of the classes themselves when selecting a final solution.

There are two other tests available in Mplus that can be used to compare models with different numbers of latent classes. The first is the Vuong-Lo-Mendell-Rubin (VLMR) test that uses a corrected distribution of likelihood differences to compare the solution with C classes to one with $C - 1$ classes (Lo, Mendell, & Rubin, 2001; Vuong, 1989). The bootstrap likelihood ratio test (BLRT) identifies the distribution of likelihood distributions by generating data from the $C - 1$ solution and running a number of models with C classes, then using that distribution to calculate the *p*-value of the difference between the model fit of the two (McLachlan & Peel, 2004). In each case, small *p*-values indicate that the C class solution fits the data better than a $C - 1$ class solution.

Two other criteria for model evaluation and selection are class homogeneity and class separation. Homogeneity is the extent to which members of the same class give the same responses (Collins & Lanza, 2010). This can be assessed by looking at the item response probabilities, such as those presented in Table 3-2. The class in column 6 of the table, characterized by the response pattern AABB, has response probabilities of 1.00, 0.97, 0.98, and 0.95. These values are so close to 1 that almost every student in that class is likely to provide the response pattern AABB. In contrast, the class in column 2 is very heterogeneous. The modal response probabilities for the bb** group are 0.62, 0.59, 0.45, and 0.44, indicating that members of the class did not give the same answers on the FMCE. It appears that Class 2 is an 'other' class comprised of students who do not fit in any of the other classes. It is common for LCA models to produce at least one small 'other' group for individuals that do not fit elsewhere. Strong homogeneity, as with the sample solution presented in Table 3-2, is desirable when selecting the best solution.

Class separation is analogous to simple structure in EFA and indicates that members of one class give responses unlike those of other classes. In other words, a response pattern with a large probability in one class has a low probability in other classes. The classes identified in Table 3-2 are characterized as AADD, bb**, aaBB, AA**, BBAA, and AABB by their modal probabilities. The AADD, BBAA, and AABB classes have very strong homogeneity, which is a necessary condition for separation. However, it is not easy to evaluate class separation from just the modal response probabilities because members of classes have non-zero probabilities of selecting off-class responses. Consider the response pattern ABDB, given by 2% of the pretest sample, which is not a part of the set of modal class descriptions. The LCA process may have combined the ABDB students with a particular class, placing them in AABB, AADD, or bb**,

or it may have generated posterior probabilities showing that they have non-trivial probability of being in any of the three. *Cross-classification* occurs when a response pattern could reasonably indicate membership in more than one class and is an adverse consequence of low separation.

Mplus provides two tools for evaluating class separation. Table 3-5 shows how individuals are classified, using the pretest Energy testlet example. Each row contains information on students who were most likely to be members of each class, while each column shows the average posterior probability of being members of each class. Classification tables with greater values on the diagonal and smaller off-diagonal values have a higher degree of separation. Separation is represented more concisely by a calculated value called *entropy*. Entropy is a signal-to-noise ratio that varies from zero (all noise) to one (perfect classification). The entropy of the sample LCA using the pretest Energy data is 0.796, just under the 0.8 value that researchers use as a benchmark for quality classification.

Table 3-5

Classification table of most likely class (row) against class membership probability (column)

		1	2	3	4	5	6
Modal Class		AADD	bb**	aaBB	AA**	BBAA	AABB
1	AADD	**0.900**	0.055	0.001	0.041	0.000	0.003
2	bb**	0.016	**0.807**	0.041	0.113	0.003	0.021
3	aaBB	0.001	0.108	**0.858**	0.029	0.003	0.002
4	AA**	0.045	0.106	0.011	**0.806**	0.014	0.018
5	BBAA	0.000	0.022	0.003	0.018	**0.957**	0.000
6	AABB	0.004	0.068	0.001	0.020	0.000	**0.906**

High levels of homogeneity and separation are desirable when selecting the best number of classes for a LCA. They also contribute to the identifiability of a latent class model. Sparse contingency tables are less problematic when cell frequencies are clustered on the table rather than spread evenly. While homogeneity and separation are desirable for estimation and

interpretation, in the extreme they make latent class modeling unnecessary. An LCA solution with perfect separation has a number of classes equal to the number of observed response patterns. Each individual has a posterior probability of one in their own class and zero in all other classes. In this case, LCA is unnecessary because the researcher can see the exact solution by looking at the raw data. The strength of latent class modeling is its ability to describe overlapping data with probabilistic responses and class membership.

Latent Transition Analysis

Latent transition analysis is a latent Markov model that acts as a longitudinal extension of LCA (Bye & Schechter, 1986; Collins & Wugalter, 1992). This type of model estimates class membership at two time points and estimates the probability of transitioning from each class to each other class. The path diagram in Figure 3-4 shows a generic LTA model, which includes arrows to represent the measurement parameters at each time point, the prevalence parameters at each time point, and the transition parameters from time one to time two. These models have many more parameters than LCAs and so tend to suffer more non-identification problems. Most researchers choose to constrain the item response parameters to be equal across time points, which not only facilitates model identification but also makes the model more interpretable.

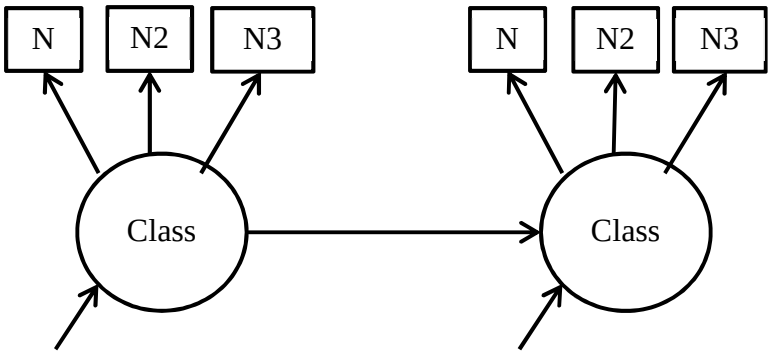


Figure 3-4. Path diagram for a generic latent transition analysis.

The prediction of posttest class membership is a multinomial logistic regression using pretest class membership as a set of independent variables. The transitions can be estimated using a probability or loglinear parameterization. Mplus uses a loglinear parameterization where each parameter represents the log-odds of transitioning from one pretest class, versus the reference class, to another posttest class, versus the reference class. There are $(C-1)(C-1)$ transition parameters, where C is the number of latent classes, and an additional set of $(C-1)$ prevalence parameters for the posttest latent variable. The prevalence parameters for the latent variable at time one are interpreted the same as in the LCA models but prevalences at time two are the means of class membership *after* accounting for the transition parameters. The LTA model can be extended to more than two time points, in which case the transitions are typically constrained to be equal from each time point to the next. This study uses only two time points to represent pretest and posttest assessments.

Constraining of item response parameters across time points makes an assumption of measurement invariance. A typical definition of invariance is that two individuals with the same level of an underlying construct should obtain the same score, despite being members of

different groups or taking the test at different times. Collins and Lanza (2010) offer a definition of measurement invariance that is specific to latent class modeling: "In LCA, an instrument fulfills measurement invariance across populations when the individuals who belong to the same latent class, but who are from different populations, have the same probability of providing any given observed response pattern." (p. 118). The first step in checking measurement invariance is to see if the two time points have the same optimal number of classes. If they do not, invariance is not necessarily violated. If a class only exists at pretest or posttest, it will not appear in the analysis of the other time point, though it is likely to appear in the solution of the LTA model. In this case, differences in classes are quantitative rather than qualitative and the LTA should have a combination of the classes at time one and time two.

In cases where the number of classes is the same, a full test of measurement invariance compares the fit of two LTA models, one with constrained measurement parameters and one with freely estimated parameters. If the less constrained model fits statistically significantly better, as determined by the likelihood ratio test, the parameters must be different across the two time points. The same process can be used to test measurement variance across known groups. Partial invariance can be tested by constraining some, but not all, of the parameters when performing the test.

If the item response parameters are significantly different across time points or groups, the models may still be interpretable. The parameters may differ statistically but provide the same substantive interpretation. This might occur when the sample size is very large, as with any significance test, or if the differences between the parameters do not change the overall interpretation of the latent classes. Collins and Lanza (2010) say that "The presence of group differences in item-response probabilities does not always rule out comparisons of latent class

prevalences, but it does mean that the comparison must be made cautiously." (p. 127) In the case of LTA, the transition parameters must also be interpreted cautiously if measurement invariance does not hold.

Multiple Group Analysis

Latent class modeling can be extended to multiple-group analysis, which is used in the present study to compare transition probabilities across treatment groups. In Mplus, multiple group latent transition analysis (mLTA) uses a categorical covariate to represent the grouping variable, similar to a MIMIC model (multiple indicators, multiple causes) in structural equation modeling. The grouping variable must be specified in Mplus using the KNOWNCLASS option, which creates an observed variable that operates as a latent variable within the software. Figure 3-5 shows the path diagram for a mLTA model, with arrows to represent each set of loglinear parameters. In addition to the parameters present in LTA, there are (G-1) group prevalence parameters, $(G-1)(C-1)$ regression parameters predicting pretest membership by group, $(G-1)(C-1)$ regression parameters predicting posttest membership by group, and $G(C-1)(C-1)$ transition parameters, where G is the number of groups. The path diagram has multiple arrows between the latent classes, indicating that separate transition parameters are estimated separately for each group (three arrows in this case because the current study analyzes three treatment groups). The measurement parameters are constrained to be equal across groups and, as with the regular LTA model, across time points. The same measurement invariance caveats apply across groups as they do across time points.

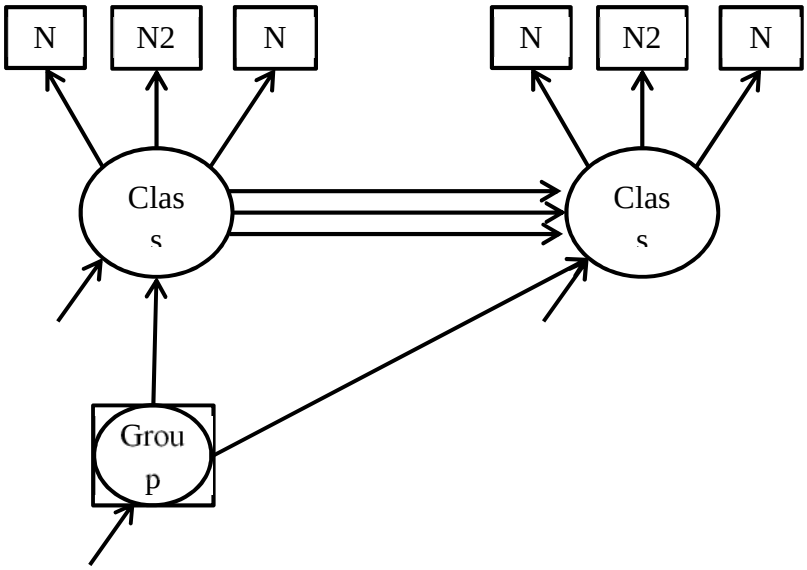


Figure 3-5. A path diagram for a generic multi-group latent transition analysis.

Collins and Lanza (2010) point out that if the item response parameters are the same across groups, then differences between the classes are quantitative and not qualitative. The goal of the current study is to see how students change their conceptions about force and motion, which assumes that the types of concepts students have are similar across institutions, treatments, and time points. Some ideas will be more common at one particular testing occasion or in one particular group, differences that can be modeled using the prevalence and transition parameters.

Measurement invariance and class separation, attributes needed for high quality classification, are a vital concern for the mLTA model used in this study. To make any inference about changes in student knowledge, information about where students begin and end the semester must be of high quality. Analogous to the impact of reliability on gain scores, uncertainty about classification at each time point compounds the uncertainty of transitions. LCA and LTA models are powerful because of their flexibility, but they are also notoriously 'fuzzy' and 'squishy.' Several authors have noted that latent class modeling is as much an art as

it is a science. Collins and Lanza (2010) repeatedly emphasize that latent class modeling requires as many judgment calls as it does statistical tests.

CHAPTER 4

METHODS

This chapter describes the specific procedures used to collect and analyze the FMCE data using latent class modeling. The first section describes the studies that produced the data sets, as well as how the data was coded and scored. Also described are the conventional scoring methods used to provide contrast to the latent class models. The next section describes the set of LCA and LTA models applied to each of the FMCE testlets. Finally, this chapter details some of the complications encountered and challenges of latent class modeling, including the Mplus error messages encountered.

Data

Data Sources and Treatment Groups

The data for this study come from two collections of Force and Motion Conceptual Evaluation results. The two studies, conducted at multiple post-secondary institutions, targeted two research-based introductory physics curricula. The data was provided by Dr. Michael Wittmann, who did not collect the data but received permission from the original researchers to share fully anonymized versions of the data. In both sets, the FMCE was given as a pretest and posttest across a semester of introductory physics instruction. Students took the FMCE voluntarily and were not graded on participation or performance.

The first data set is a collection of FMCE results from classrooms using the Tutorials in Introductory Physics curriculum (McDermott & Shaffer, 2001). This is referred to as the TUTORIAL group. The data came from eight sections of introductory college physics courses at multiple four-year universities. The Tutorials in Introductory Physics curriculum uses paper-

and-pencil activities that students complete in small groups. The tutorials can be added to traditional instruction or can be adopted as a part of a reform-based curriculum.

The tutorial activities are based on extensive research on student misconceptions, most of which occurred at the University of Washington in the 1980s and 1990s (e.g. McDermott, Rosenquist, & van Zee, 1987). The activities themselves use lines of questioning that lead students to confront their own misconceptions. The aim of each activity is to have students logically prove the correct concept and then activate their own incorrect intuitions. The final questions of each activity ask students to reconcile the two, contradictory ideas. As an example, the Reverse Direction concept is addressed by having students draw 'change-in-velocity vectors' for a ball rolling up, and back down, a ramp. The step-by-step questioning leads student to recognize that, while the velocity is different at each point on the ramp, the change-in-velocity vector is the same length at each point, including at the top of the ball's motion. The students are then asked to reconcile this result with their initial predictions which, more than likely, say the force on the ball is positive, zero, and negative.

The second data set came from a large scale evaluation of the Real-Time Physics with Interactive Lecture Demonstrations curriculum (RTP/ILD; Sokoloff, et al, 2011). The evaluation took place in 1998-2002 across six institutions that included a land grant state university, a community college, and a military academy. FMCE data was collected from 65 sections of introductory college physics courses. Dr. Wittmann performed the final analyses of the evaluation data but the results have not been published.

The Real-Time Physics curriculum uses sensor technology and graphing software to engage students in the process of collecting and interpreting data. The small-group activities in the RTP curriculum can take place during lectures or take the place of laboratory or recitation

sections. The RTP activities address misconceptions by presenting students with data that they collect themselves in key situations. For example, RTP addresses the Reverse Direction context by having students toss a ball up into the air above a motion sensor. The sensor feeds data to a computer, where the students can create position, velocity, or acceleration vs. time graphs. They can see for themselves that, while the velocity of the object passes through zero as it changes direction, the acceleration remains constant. The Interactive Lecture Demonstrations are an optional supplement to the RTP materials. They are similar activities, using sensors and computer software, which instructors perform during lectures for full class instruction and that also involve student participation.

The RTP/ILD evaluation was plagued by methodological and implementation issues, two of which are relevant to data preparation. First, the evaluators used the Force Concept Inventory during the first year of the evaluation rather than the Force and Motion Conceptual Evaluation. The first year of the study was when they collected baseline data from classrooms using traditional pedagogy. As a result, only three of 65 sections in the RTP/ILD study provided control data. Second, the intervention suffered from inconsistent fidelity of implementation, where instructors used varying amounts of the RTP and ILD curricula. Fortunately, it was possible to use the fidelity information to create an ad hoc control group. Dr. Wittmann was able to provide implementation notes taken by RTP/ILD evaluators and instructors. According to the notes, 42 sections used both parts of the RTP/ILD curriculum, 20 sections used *either* Real-Time Physics or Interactive Lecture Demonstrations, and three used neither. Using this information, the data were split into two groups, consisting of students who received both parts of the intervention and students who received one part or neither parts.

The two groups are referred to as the BOTH and NOT BOTH groups rather than using the labels of treatment and control. This division would be an inadequate attempt to create a counterfactual group in a study aimed at testing the effectiveness of the RTP/ILD curriculum. However, the goal of the current study is to evaluate multi-group latent transition analysis as a tool for detecting differences in group transitions. To that end, it is only necessary that the groups are meaningfully distinct. The BOTH and NOT BOTH labels highlight that the students received different instruction without asserting that observed differences are intentional and due to the impact of the curriculum. Table 4-1 shows that the groups were sufficiently sized to perform the LTA analyses, with approximately 1500 students in BOTH and 1000 students in NOT BOTH.

Table 4-1

Numbers of students in reported data that received RTP and/or ILD instruction

	No ILD	ILD	Total
No RTP	336	179	515
RTP	451	1541	1992
Total	787	1711	2507

Note that while the TUTORIAL group received the full tutorial instruction and the BOTH group received full RTP/ILD instruction, the NOT BOTH group is a heterogeneous set of students who received a variety of types of instruction. Some students in the NOT BOTH group were taught with deliberately traditional methods while other students in the group received either RTP or ILD. Some NOT BOTH students were taught by instructors who intended to use RTP/ILD but their commitment to the program was so poor that the instruction could not be considered RTP/ILD. The data used in this study came from a diverse set of institutions, so the

physics courses in *all* treatment groups must have varied in structure and in quality. However, the heterogeneity within the NOT BOTH group is vast. The NOT BOTH label is intended to reflect the heterogeneity and the only true statement that can be made about the group is that students took a physics course but did not receive the full RTP/ILD curriculum.

Data Cleaning and Coding

The sets of FMCE data are incomplete for a number of reasons. College students are likely to add, drop, or retake courses. Students that do remain in physics courses from beginning to end might still miss either the first or last class of the semester. A small percentage of students may have chosen not to complete the assessment. Other students may not have been able to complete the FMCE in the allotted time, or may have skipped entire sections of the test. The completeness issue is compounded by varying degrees of pre-cleaning. Some of the participating institutions provided all data, no matter how incomplete; others institutions only submitted data from students who fully completed both the pretest and posttest. Given the issues with data collection and data cleaning, the data should not be considered a random or representative sample. The samples are broad, including a variety of students, but may suffer from sampling bias due to missing, relevant groups of students. As with the design issues described above, bias and representativeness do not threaten my primary research objective, which is to evaluate multi-group latent transition analysis as a methodological tool.

The data, which came from several institutions using different data formats and cleaning procedures, needed to be cleaned and standardized. First, all illegitimate responses (any character not available as a response for that particular item) were coded as missing. Then pretest and posttest results were matched by student and redundant student data were deleted. Next, three variables were created for classifying data completeness for each student. The first

variable is a count of number of missing responses on the FMCE. The second is a dichotomous variable to indicate whether a survey was missing entirely. The third variable indicates whether the student completed the FMCE. Surveys with 47 missing items were marked as 'missing,' while surveys with three or fewer missing responses were marked as 'complete.' The remaining surveys, with 4 – 47 missing items, were labeled according to a particular heuristic. If the student responded to at least one item on each testlet, the survey was marked 'complete.' If any one testlet was missing all responses, the survey was marked 'incomplete.'

The rationale behind this heuristic is that it is impossible to know if a student deliberately refused to answer a testlet or if they missed it entirely. Some FMCE administrators use two-sided photocopies that put entire testlets on the back sides of sheets of paper, making it easy for students to miss entire testlets. Most of the incomplete tests were missing answers to the Energy testlet, and it is impossible to know if the student ran out of time, chose not to answer the questions, or if they missed the testlet on the back side of the final page. It seems reasonable to calculate raw scores for students when there is evidence that they at least saw the testlet.

The count data in Table 4-2 shows the number of missing, incomplete, and complete surveys in all three treatment groups. Table 4-3 gives tallies for the number of students reported in each treatment group and the number of matched pretest-posttest pairs of complete FMCEs. Relatively few surveys were marked as incomplete. The TUTORIAL group had many more missing and incomplete surveys than the other two groups. It is possible that this difference is due to some institutional differences in attendance or survey administration. However, some RTP/ILD institutions cleaned the data heavily before submitting it to the original evaluators while the TUTORIAL group was left in its original state. So the proportion of missing and

incomplete data in the TUTORIAL group might not be abnormal. It might be typical of uncleaned college survey data.

Table 4-2

Number of entirely missing, incomplete, and complete surveys in each treatment group

	Pretest			Posttest		
	Missing	Incomplete	Complete	Missing	Incomplete	Complete
NOT-BOTH	59	23	884	103	9	854
BOTH	38	39	1464	102	3	1436
TUTORIAL	246	109	1707	748	14	1300

Table 4-3

Number of students in each treatment group and the number of matching pretest-posttest FMCE pairs

	Total Number of Students	Matched Pairs of Complete Surveys
NOT-BOTH	966	776
BOTH	1541	1360
TUTORIAL	2062	995

Among the matched pairs of complete pretests and posttests, the proportion of missing data was less than 0.5%. This small proportion of missing data should not pose any threat to the validity of the study, and is likely a much smaller than the other sources of bias described above. For conventional scoring, missing responses were counted as incorrect. Latent class modeling in Mplus uses all available information and is not impeded by a small proportion of missing data.

The next problem to address was that of entirely unselected responses, those that no students in the sample selected. Mplus does not have an option for specifying the number of

levels of a categorical variable. It uses the data file itself to determine the correct number of levels. For example, Item 2 has eight response options, A through G and J, so Mplus would see eight distinct values in the Item 2 column and assume the nominal variable has eight categories. However, no students in the posttest data selected G so Mplus assumed the variable only had seven levels. The same problem occurred with the response J for Item 13 at posttest. Having the incorrect number of levels of a variable is problematic when constraining item response parameters in latent class models. To address the issue, fictional students were created and added to the data files for latent class modeling (they were not included in any conventional analyses). Each fictional student gave the response that no other student gave, and all other responses were marked as missing data.

Scoring

Conventional analyses were performed to characterize the data and to test the assertion that categorical scoring is more useful for analyzing the FMCE than conventional methods. To this end, the FMCE was scored with a template using recommendations from Thornton and Sokoloff (1998) and Smith and Wittmann (2008). The designers of the FMCE recommend that seven items (5, 6, 15, 33, 35, 37, 39) be omitted from scoring. Some of these items are extremely easy and intended as red flags for students who cannot read English or choose to guess on all items, while other items were found to be confusing for students. None of these items were scored in any analyses.

The template uses Wittmann's recommendation to split the FMCE items into seven testlets by the physics concepts that they target (CITE THE TEMPLATE). Table 4-4 names the testlets and shows which items are assigned to each. There are other defensible ways to score the FMCE, perhaps splitting the Newton's Third Law items into 'collisions' and 'pushing' or the

Reverse Direction items into ‘car’ and ‘coin.’ The division of items used in this study emphasizes the conceptual focus of the testlets. The scoring template also includes scoring modifications to avoid false positives. It is possible for students to answer the Reverse Direction ‘falling object’ items correctly without demonstrating mastery of the main concept. Thornton and Sokoloff (1998) recommended that each group of three RD items be worth two raw score points if, and only if, all three items receive a correct response. Students who do not get all three items correct receive zero points. Similarly, Smith and Wittmann (2008) suggest that item 32 is sometimes a false positive and should only be considered correct if items 30 and 31 are also correct.

Table 4-4

Scoring template for the Force and Motion Conceptual Evaluation by testlet

Name	Abbrv.	Items	Modifications	Points
Force Sled	FS	1-4, 7		5
Reverse Direction	RD	8-10, 11-13, 27-29	Two points assigned for each group of 3 items if all correct	6
Force Graphs	FG	14, 16-21		7
Acceleration Graphs	AG	22-26		5
Newton 3	N3	30-32, 34, 36, 38	Three points assigned if 30-32 all correct	6
Velocity Graphs	VG	40-43		4
Energy	E	44-47		4

The Energy testlet is considered an optional section of the FMCE. Its items assess a concept that is somewhat removed from the rest of the instrument, asking students about energy and motion, rather than force and motion. Many practitioners choose to include the Energy

testlet, as it supports the overall purpose of assessing student conceptual knowledge in physics. Since the Energy testlet is well suited for latent class analysis, it was also included in the analogous raw score results. With the Energy testlet, the maximum total score of the FMCE is 37 points.

Means and standard deviations at pretest and posttest were computed for each of the three groups of students. Cronbach's alpha was computed at pretest, at posttest and for the combined pretest-posttest data. Analysis of covariance procedures tested for differences in mean total scores between groups at posttest, controlling for pretest. In addition, a number of psychometric techniques including factor analyses and item response theory analyses were used to explore the properties of the FMCE. The factor analysis and item response theory results are tangential to the current study and will not be discussed in depth in this dissertation, though they are presented in Appendix B.

Latent Class Modeling

The Modeling Process and Model Interpretation

The primary goal of this study was a multi-group latent transition analysis to test whether transition parameters differ across groups. The modeling process began with pretest, posttest, and combined data LCAs to obtain preliminary information about class structure. Next, a single group LTA informed the decisions made for the multi-group latent transition analysis. Finally, the mLTA model was fitted to the data and compared against a constrained model with equal transition parameters across groups. The same modeling process (LCA → LTA → mLTA → constrained) was applied separately to each testlet. The Velocity Graphs testlet was excluded from the latent class modeling process because the items, which are too easy for college-level students, provide very little information about student mental models.

Throughout these analyses, the items were entered as nominal variables where each categorical level corresponded to one item response (A, B, C, and so on). It is possible to model the items, or even the latent class variables, as ordinal but the current study makes no assumptions about the correctness of various mental models or responses. While it is true that each latent variable has one correct class and each item has one correct response, it is not possible to put the classes into an order of correctness or maturity. The LTA results presented in Chapter 5 may be the first steps in identifying ordinal structures from empirical results. Future research may examine the stages of learning physics concept, leading eventually to statistical models that use ordinal variables. The models in this study, those with only nominal variables, belong to the family of multinomial logistic regression models.

Mplus estimates latent class models using a maximum likelihood estimation procedure with robust standard errors that takes into account all available information. The LCA models were estimated on a single-processor laptop computer while the LTA and mLTA models were run on the High Performance Computing (HPC) cluster at the University of Connecticut's Booth Engineering Center for Advanced Technology. The HPC cluster provided 24 processing cores, running Mplus Version 7.3with parallel processing to greatly reduce computation time. Mplus uses a default of 500 iterations of the maximum likelihood procedure to achieve convergence. The number was sufficient to converge on the LCA model solutions but was increased to 1000 to for the LTA and mLTA models. In each analysis, Mplus input code specified 200 replication starts and 20 replication finishes. As described in Chapter 3, increasing the number of replications improves the chance of finding the solution with the global maximum of the log likelihood. With the exception of the Reverse Direction testlet, which had a number of

convergence and estimation problems, all models had multiple replications that converged on the same best solution.

A number of criteria were considered when selecting the best number of latent classes for each LCA, LTA, and mLTA model. Unfortunately, many of the criteria described in Chapter 3 provided no useful information. The VLMR test always showed a significant difference between adjacent numbers of classes while the BLRT never showed a significant difference. Muthen and Muthen (2009a) say that the VLMR and BLRT sometimes fail to discriminate between solutions. He suggests that BIC and class structure should be the most important criteria and that VLMR and BLRT should be used to adjudicate when the situation is unclear. Unfortunately, the class structures provided little information that was useful for choosing a most appropriate number of classes.

Two adjacent solutions of the same testlet data, with C and $C - 1$ classes, appeared very similar on the surface. The idea of *surface similarity* will continue to appear throughout this dissertation and is not an official term used by LCA researchers, so it is important to provide an operational definition. Latent classes are often described in terms of their modal responses, the values that members of the class are most likely to generate. Those modal response labels can look identical from solution to solution, though the models are somewhat different. In some cases, increasing the number of classes by one has little effect on the class structure; it simply removes a small group of 'oddball' individuals and gives them their own class. The solutions appear to be 'similar on the surface' because the interpretations of the most populated classes are identical.

In other cases, there may be parameters that change as the model changes but those parameters do not describe the responses that are most often selected. The solutions appear

similar to visual inspection and are interpreted similarly, though other parameters may differ significantly. These 'off modal' parameters describe how other, less frequent response patterns may be folded into a larger class. On the surface, the classes appear similar because the classes have the same modal response, though some response patterns may be sorted differently. The term surface similarity is used to describe solutions that appear similar to interpretation, though they are likely to be statistically significantly different. To some extent, surface similarity is a limitation of using modal parameters for interpretation, but the models have far too many parameters to interpret by inspection, so modal responses are presented in Chapter 5.

In comparing models with different numbers of solutions, the largest and most important classes tend to remain the same while the smaller classes tend to split and sometimes recombine. The LCA solutions for most testlets included one class for the correct answer, one class for the common misconception, one for a hybrid or dual conception, and a set of 'other' classes. For example, every Force Sled model produced a correct BDFFB class, an incorrect ABCGE class, and a hybrid ABFGB class. Only the 'other' classes differed across the five, six, and seven class solutions. Increasing the number of latent classes causes individuals to 'split off' from larger classes, forming increasingly smaller and less relevant subgroups.

Table 4-5 provides an example taken from the pretest Force Sled data where the addition of a seventh class had very little effect on the overall architecture of the classes. Both solutions include the three main classes but the seven-class solution also included a class labeled AB*Ga. This 'other' class represents a very small proportion of the sample. A reduction in the size of the AB*G* class hints that the AB*Ga students have split off from the AB*G* students. On the surface, it appears that the difference between those that split off and those that remained in the same class is that they were more likely to select response A on the last item. However the non-

modal parameters reveal another difference between the groups. The AB*Ga students were split on the third item, 40% selecting E and 40% selecting G, while the AB*G* students had near zero probabilities of selecting E or G on the third item. The classes might also be labeled as **AB[E or F]Ga** and **AB[neither E or F]G*** but, while this is a more complete description, the notation is too cumbersome to report in all cases.

Table 4-5

Modal item responses for two solutions of the pretest Force Sled LCA

Six Class	Seven Class
BDFFB	BDFFB
ABCGE	ABCGE
ABFGB	ABFGB
AB*G*	AB*G*
B****	B****
gFeAc	gFeAc
	AB*Ga
BIC: 33486	BIC: 33499

In this dissertation, when I refer to solutions as ‘similar on the surface’ I mean that the modal response patterns of the classes are the same and, if the number of classes differs, differences appear to be a shuffling of students among the small ‘other’ classes. Surface similarity made it difficult to use class structure to select the best number of latent classes because adjacent (C versus C + 1 class solutions) always appeared similar on the surface. Looking at the list of classes in Table 4-5, for example, I cannot justify why one solution would be superior to the other. The AB*Ga class does not seem to provide much meaningful information, so it seems prudent to err on the side of parsimony and choose the six-class solution.

In selecting a number of latent classes, I also considered the entropy value of each solution. While entropy is not a measure of model fit, it does give an indication of whether the solution is useful. The rule of thumb in latent class modeling is that the entropy should be greater than 0.8 (Muthen & Muthen, 2009a). Unfortunately, entropy was not very useful in selecting the best number of latent classes because all models had entropy values of 0.8 or more. The entropy varied across adjacent solutions, but did not change dramatically. Given that entropy, substantive reasoning, VLMR, and BLRT were not very helpful, BIC became the primary criterion for choosing the best number of classes.

The Latent Class Models

The exploration of each testlet began with latent class analyses of the pretest, posttest and combined data, as in Figure 4-1. The purpose of these analyses was to identify the class structure at each time point, evaluating measurement invariance across times and informing the LTA and mLTA models. If classes at pretest and posttest were meaningfully different, then constraining measurement parameters to be equal during the LTA would be problematic. On the other hand, if some classes were simply present at one time and not the other, the latent transition models should function normally. In that case, the LTA identifies the one-time-only classes but assigns very few students to those classes at the unpopulated time point. In most testlets, the best number of classes found in the pretest data was greater than in the posttest data. There may be a meaningful difference between pretest and posttest students that leads to fewer latent classes but the matter requires further qualitative research.

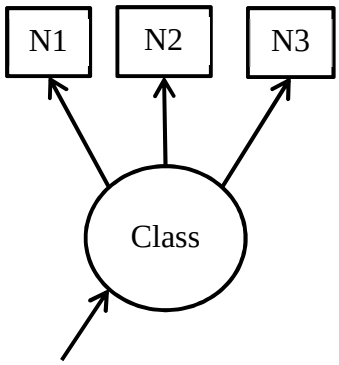


Figure 4-1. Path diagram for a single Latent Class Analysis

In all cases, the best number of classes for the combined data was larger than the combination of the number of pretest and posttest classes. Table 4-6 shows the best solutions of the LCAs of the Force Sled testlet as an example. The pretest, posttest, and combined data had six, four, and eight solutions, respectively. It appears that, when the two data sets were combined, certain response patterns reached a 'critical mass' where they needed to be identified as separate classes in order to maximize model fit. In this case, the main classes are the same across all three models, and the combined data contains the classes represented in the pretest and posttest solutions with some extra 'other' classes. The AB*G* of the pretest data appears to be split into ABgGA and ABeG* in the combined data, presumably because the additional cases in the posttest data provided enough information to split the students. Similarly, the B**** and Bd*f* students in the pretest and posttest data recombined to form two groups in the combined solutions: B*cfe and B**F*.

Table 4-6

Hypothetical class structure, illustrating the pattern of classes when two data sets are combined

Pre LCA	%	Post LCA	%	Combined	%
BDFFB	13	BDFFB	41	BDFFB	24
ABCGE	48	ABCGe	28	ABCGE	37
ABFGB	18	ABfGb	23	ABFGB	17
AB*G*	14			ABgGA	6
B****	5	Bd*f*	8	B*cfe	4
gFeAc	1			B**F*	4
				ABeG*	3
				*****	1

The next model applied to each testlet was a latent transition analysis, shown graphically in *Figure 4-2*. The item response parameters at pretest and at posttest were constrained to be equal so that the latent classes would be forced to have the same interpretation at the two time points. The best fitting LTA models had the same number of classes as the combined data LCAs and had very similar class structures. LTA analysis in Mplus uses all available data, so students who only provided pretests or posttests were included. The information provided by pretest-only and posttest-only students was used in estimating the item response parameters and so contributed to the characterization of the latent classes. These students' data did not contribute to the estimation of the transition parameters.

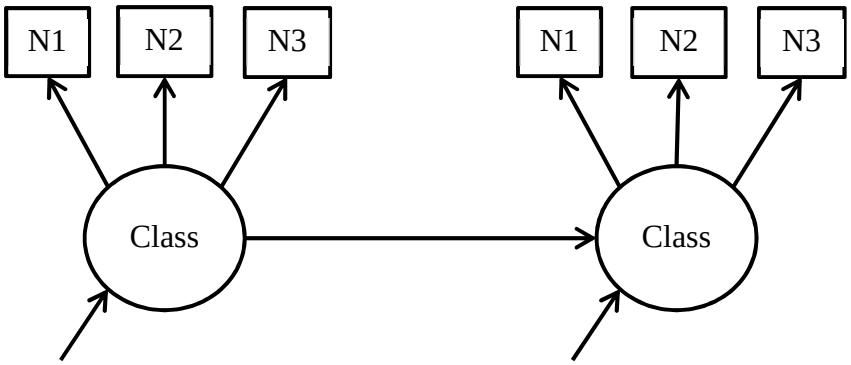


Figure 4-2. Path diagram for a latent transition analysis.

Following the LTA, the mLTA model generated transition parameters for each of the treatment groups, structured as shown in Figure 4-3. This model used the KNOWNCLASS option of Mplus as described in Chapter 3 to include the treatment group variable in the model. The grouping variable predicts the pretest class, and is used to create three separate sets of transition parameters from pretest to posttest. The model has one set of prevalence parameters at each of pretest and posttest, three sets of transition parameters, two sets of item response parameters that were constrained to be equal, and a set of parameters predicting pretest and posttest by group membership.

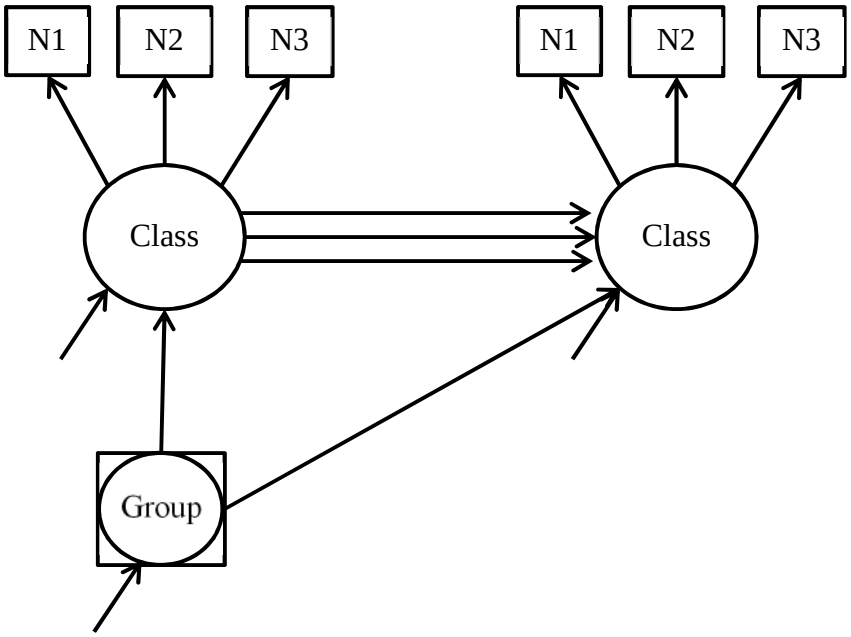


Figure 4-3. A path diagram for a multi-group latent transition analysis using the Mplus knownclass option to create a MIMIC model.

Running several mLTA models with different numbers of classes consistently produced best fitting solutions with fewer than the latent transition analyses. In the case of Force Sled and Acceleration Graphs, the best mLTA had *two* fewer classes than the best LTA model. There are two possible explanations for this phenomenon. The first is that the mLTA models increased in the number of parameters faster than the LTA models. BIC increases linearly with the number of parameters in a model, so if the mLTA parameters increased more with each class, the BIC increased faster, hitting a minimum at a lower number of classes. The other explanation approaches the issue from the context of information and classification. Adding a group variable provides information that was not previously available, information that is exploited by the latent class estimation process. The extra information allows the model to describe the sample with fewer classes. The class structures of the LTA and mLTA were similar on the surface and in some cases the mLTA solutions seemed 'cleaner.'

The central analysis of this study was to test whether the transition parameters were statistically significantly different across treatment groups. If so, then the mLTA model is appropriate for detecting differences in conceptual change across groups. To show this, the model in Figure 4-3 needed to fit the data significantly better than one where the transitions were constrained to be equal. The LTA model shown in Figure 4-2 is not nested within the mLTA model, because of the KNOWNCLASS variable, so the LTA cannot act as the constrained constrained model. Instead, the model in Figure 4-4 was used because it includes the grouping variable (as a predictor of pretest class membership, as in the mLTA), but uses a single set of transition parameters. The effect of group on posttest class was removed to prevent the transition differences from 'sneaking into' the model through other parameters.

To compare the fit of the models in Figures 4-4 and 4-3, there are two options. Muthen and Muthen (2009b) recommend using the likelihood ratio test, where $2(LL_2 - LL_1)$ is distributed as a chi-square. This test is recommended whenever models are nested, as is the case with the multigroup LTA and its constrained counterpart. If significant, the additional parameters in the model that generated LL_2 made it fit the data significantly better than the more parsimonious model. The other option is to compare the BIC values for the two models, where a lower BIC indicates better fit. Collins and Lanza (2010) presented an example where the likelihood ratio test and the BIC disagreed. The multigroup LTA fit statistically significantly better than the constrained LTA by the likelihood ratio test, though the constrained model had a lower BIC. The authors did not offer any advice on how to reconcile the conflicting signals, nor have other authors in the body of literature on comparing latent class models. It remains unclear how to interpret conflicting indicators.

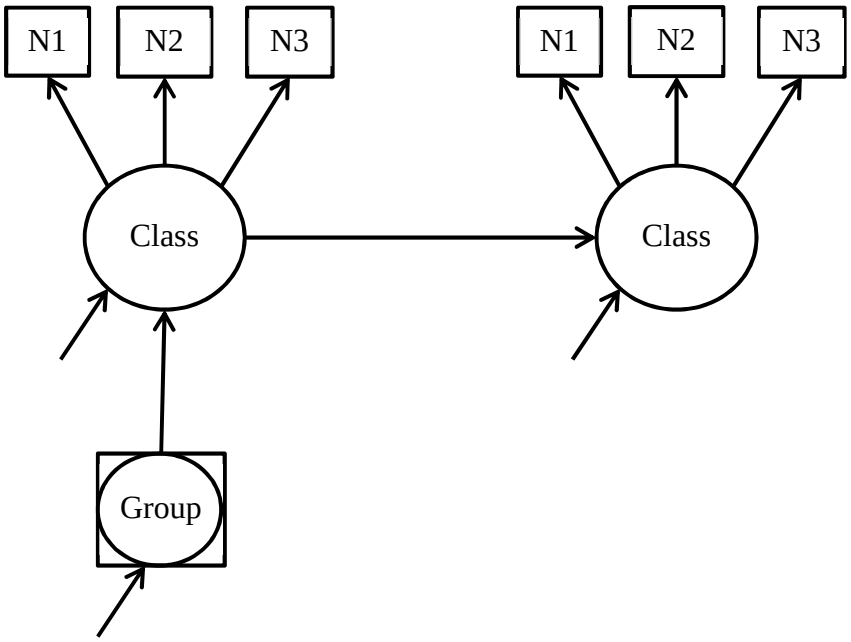


Figure 4-4. The comparison model for testing the statistical significance of differences among transition parameters.

Modeling Issues and Complications

Latent class modeling is often described as being 'just as much art as science' because of the large number of decisions with no objectively correct choice. Beyond the 'fuzziness' that is often encountered, there were several complications—ranging from inconveniences to severe limitations in study validity—encountered during the modeling process. First, Latent class solutions often include item response parameter estimates that are arbitrarily large or arbitrarily small. This is particularly true in situations where all variables are nominal. Mplus constrains these values to 15 or -15 on a logit scale, values that indicate extremely large and extremely small odds, respectively. The warning list that Mplus generates for each analysis may intimidate researchers into believing that something is wrong. The reality is that extreme values are often indicators of homogeneity. If a class has a measurement parameter fixed at 15 for a particular response, then all members of that class gave that response. Likewise, a measurement parameter set at -15 indicates that no member of the class gave that particular response. A solution with

many extreme values acts as a strict set of rules for classifying students. In the LTA or mLTA models, some transition parameters were fixed at -15, indicating that no students in the sample made a particular transition across classes. Abar and Loken (2012) conducted a simulation study with intentionally unidentified models and found that Mplus tended to identify them anyway, using these constraints.

As mentioned in Chapter 3, there are two possible parameterizations of the structural part of the latent transition models, probability and loglinear parameterization. Mplus defaults to loglinear but offers probability parameterization as an option. Unfortunately, no models converged on a solution when using the probability parameterization. This is unfortunate because that option makes it easier to compare structural (transition) parameters across groups. The loglinear transitions are not directly comparable because they are composites of multiple loglinear parameters that refer to different reference groups.

The next issue is one of model identification. All models described in the previous section have a positive number of degrees of freedom, which is a necessary but not sufficient condition for model identification. Latent class models can suffer identification problems due to the features of a particular data set and particular model. Situations with many latent classes or sparse contingency tables are more likely to have identification issues (Collins & Lanza, 2010). The FMCE testlets generate extremely sparse tables. The Energy testlet is the smallest table with four items each with five responses. There are $5^4 = 625$ possible response patterns. The sample of approximately 3500 students appropriately fills the Energy testlet table, meeting the rule-of-thumb guideline of N being five times larger than the number of cells in the table. The sample fails to meet the same guideline for the Force Graphs testlet which has $9^7 = 4{,}782{,}969$. Fortunately, the sparseness of the FMCE contingency tables is counteracted by a high degree of

class separation, as indicated by the greater than 0.8 entropy of the model solutions. Identification issues can also occur when there are not enough individuals in a particular latent class to determine parameter values (Muthen, 2007). This may be the cause of the problems with the Reverse Direction testlet, described in Chapter 5. Model identification problems may ease if the number of classes is reduced, increasing the number of individuals in each class.

In some cases, Mplus pinpointed specific parameters that could not be identified with the information given. Often these parameters were for response options very rarely selected, or for very rare transitions between classes, in the smallest of the latent classes. One possible method for working around these unidentified parameters is to fix them to zero probability (-15 in logit units). Unfortunately, due to class label switching, fixed parameters get assigned to different classes in each run. In theory, label switching can be controlled with the use of starting values for measurement parameters but such attempts with this data did not work. Moreover, it is not clear that fixing the unidentified parameters to -15 would be appropriate. If the offending parameters describe unlikely responses or transitions, then the few cases of those responses or transitions that appear in the data would likely be associated with the smallest classes. It is the small 'other' classes that become umbrellas for odd responses. It is possible that the unidentified parameters should be set to zero probability for all of the major classes and allowed to be small-but-non-zero for the other class. Unfortunately, that is not a viable strategy for increasing identifiability.

Mplus sometimes produces error messages to say that certain parameters have been fixed because of a non-positive definite derivative product matrix. These error messages can be indicators of non-identification (Bengt Muthen, 2007). However, Bengt Muthen explains that

arbitrarily large thresholds can generate these error messages even in models that are fully identified:

> "Thresholds that go large like that are harmless causes of the non-pos def message. With large thresholds, the information matrix estimate obtained by the first-order derivative approach can be numerically determined as singular. Degree of singularity is measured by the condition number, which is the ratio of the smallest to largest eigenvalue of the info matrix estimate. You don't want a very small condition number and 0.146D-12 is very small, very close to exactly zero in machine numerical precision terms." (Muthen, 2006)

When I encountered the non-positive definite matrix error (NPD) I recorded the value of the condition number and assumed that values smaller than 10^{-10} indicated non-identified models. These models still converged and produced interpretable results but certain parameter values were fixed at extreme values which may have affected other aspects of the model. The results presented in Chapter 5 include error messages and condition numbers. Models with identification issues are singled out in the text.

In latent class models with nominal variables, Mplus frequently produces error messages explaining that a multinomial logit parameter needed to be fixed during estimation to avoid a singularity of the information matrix. This error refers to the procedure that Mplus uses to generate standard errors, which are not generated as a part of the estimation process. Linda Muthen explains in the Mplus help forums that the error message is likely not a cause for concern. “It means most often that some classes do not have variation in some covariates so regression coefficients cannot be determined. That is ok and often good in that it means that classes are clearly different [with respect] to the covariate.” (Muthen, 2007). It is theoretically possible that these errors are produced as a result of non-identification, but are more likely a lack of variation in a posttest class with respect to a specific pretest class. The FMCE models

generated a large number of these information matrix (IM) errors. They manifest in the transition tables in Chapter 5 as transition probabilities set to zero.

The final issue to discuss is that of measurement invariance. In latent class modeling, measurement invariance conceptually refers to whether the definitions of the classes are the same across time points or across groups. Classes are defined by their item response parameters, so the statistical definition of measurement invariance is whether those parameters are statistically significantly different across groups. The tests for invariance are described in Chapter 3, including a comparison of the best number of classes and the statistical fit comparison of constrained vs. unconstrained models. Invariance tests across time points and treatment groups encountered two major problems. The first is that the unconstrained models, with differing measurement parameters across time points or groups, would not converge or were unidentified. The second problem is that the model solutions, those that provided any interpretable results, appeared similar on the surface. Whether statistically significantly better fitting or not, the classes *appear* similar. The border between invariant and non-invariant is not clear (Collins & Lanza, 2010).

As described in previous sections, testlets did not have the same best number of classes at pretest and posttest, a fairly clear violation of measurement invariance. At the same time, the classes in the combined LCA and LTA models tended to make sense—a combination of the pretest and posttest classes with a few extra 'other' classes. It is not clear if the FMCE class structures are problematically different across time points. Much of the difference in class structure can be represented by the proportion of students in each class at pretest and posttest. For example, the extra posttest class might be interpretable but so few students gave the particular response pattern at pretest that the class did not appear in the pretest LCA. The

measurement parameters were constrained across time points, assuming invariance that may not be true. The classes seem similar overall, but future research should be aimed at how pretest and posttest LCA results compare.

Measurement invariance tests across treatment groups were similarly concerning. Tests of invariance involved only pretest or only posttest data, constrained across groups and freed across groups. The free parameter models all fit better than the constrained models; as expected, some fit statistically significantly better. However, the freely estimated models consistently generated the NPD errors described above, with very small condition numbers, indicating problems with model identification. It is not surprising that the models would have identification problems, since the unconstrained models have nearly three times as many parameters as the constrained ones and fewer students in each class. The sample of 3500 students is sufficient for estimating a model with 200 free parameters but struggles with nearly 600 parameters. Because of these non-identification warnings, it is difficult to make any claims about measurement invariance.

The LCA models with freely estimated measurement parameters converged on solutions, likely because Mplus automatically constrains parameters to navigate through the NPD and IM errors. I examined the solutions of each of the three groups and found that the classes were similar on the surface. In some cases, the unconstrained model fit significantly better, indicating that the item response parameters should not be constrained, but even in these models the classes appeared to be similar. Each group had a common misconception class, a Newtonian class, and a hybrid concept class, along with similar looking ‘other’ classes. It is possible that the parameters are different enough to generate significantly different model fit but not different enough to change the substantive interpretation of the classes. On the other hand, it is possible

that the classes do have some substantial differences that are not readily apparent. As with the invariance over time, I cannot make any strong claims about invariance over treatment groups.

CHAPTER 5

RESULTS

This chapter describes the outcomes of the analyses described in Chapter 4. It begins with the descriptive statistics and mean score comparisons that would accompany a typical controlled trial study. The sections that follow describe the results of the latent class modeling of each testlet. The descriptions include estimation histories, most popular response patterns in the data, modal response descriptions of latent classes, LTA transition parameters, and multi-group transition parameters. Some interesting effects are noted in the text while the broad interpretations and conclusions are presented in Chapter 6.

Descriptive Statistics

Table 5-1, Figure 5-1, and Figure 5-2 characterize the FMCE total scores with means, standard deviations, pretest score distributions, and posttest score distributions. Each figure uses all complete pretests or complete pretests, not just those from the set of matched pairs. The pretest scores show strong positive skew, indicating a difficult test. Note that the four items of the velocity graphs section were extremely easy and were answered correctly by almost all students. Given this result, a score of 4 was effectively a 0, and hence the distributions of pretest scores show a strong floor effect. The pretest means are statistically significantly different across treatment groups (F (2, 4052) = 10, p <.001), though visual comparison shows that the distributions are similar.

Table 5-1

Means and standard deviations of total FMCE scores by treatment group

Group	Pretest		Posttest		
	N	Mean (SD)	N	Mean (SD)	Effect Size (d)
Not-Both	884	8.6 (7.2)	854	16.7 (10.3)	0.91
Both	1464	9.3 (7.5)	1436	22.9 (10.4)	1.50
Tutorial	1707	10.0 (8.1)	1300	21.9 (10.7)	1.25

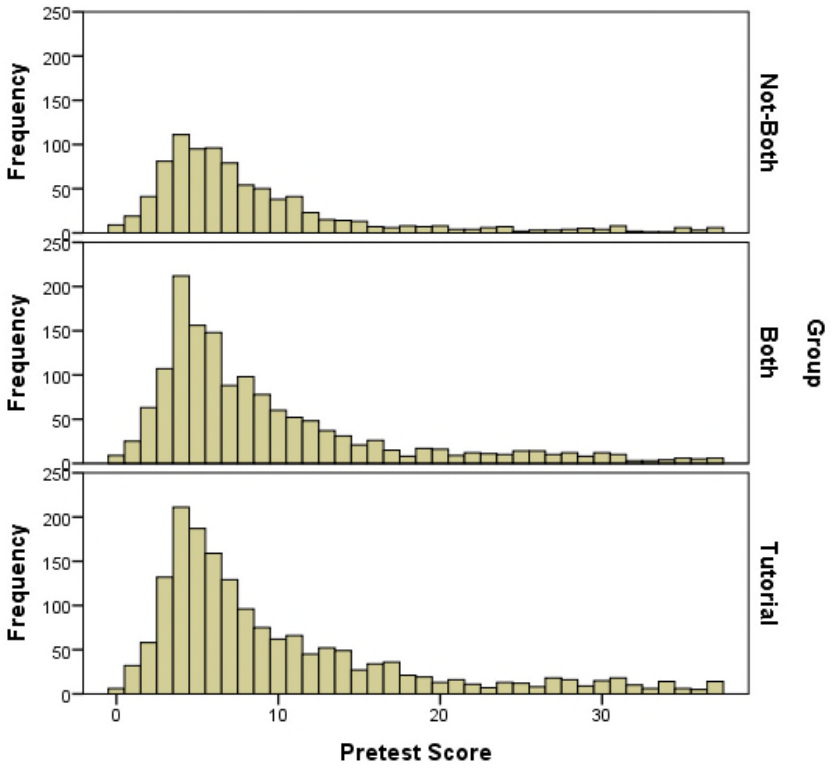


Figure 5-1. Distribution of total FMCE score at pretest by treatment group.

The posttest scores in Figure 5-2 appear to be almost uniformly distributed across the range of scores. The distributions show evidence of a minor ceiling effect in the BOTH and TUTORIAL groups. The posttest score distributions seem noticeably different across groups, with statistically significant differences among the mean scores ($F(3, 3589) = 100$, $p < .001$). The uniform score distributions may seem unusual in educational research, where normal distributions are the norm, but physics education researchers would likely not be surprised by these results. At the end of a semester of instruction, many students 'get it' and for them the FMCE is an easy task. For many other students, misconceptions persist through the semester despite the learning to solve physics problems and accumulating declarative physics knowledge.

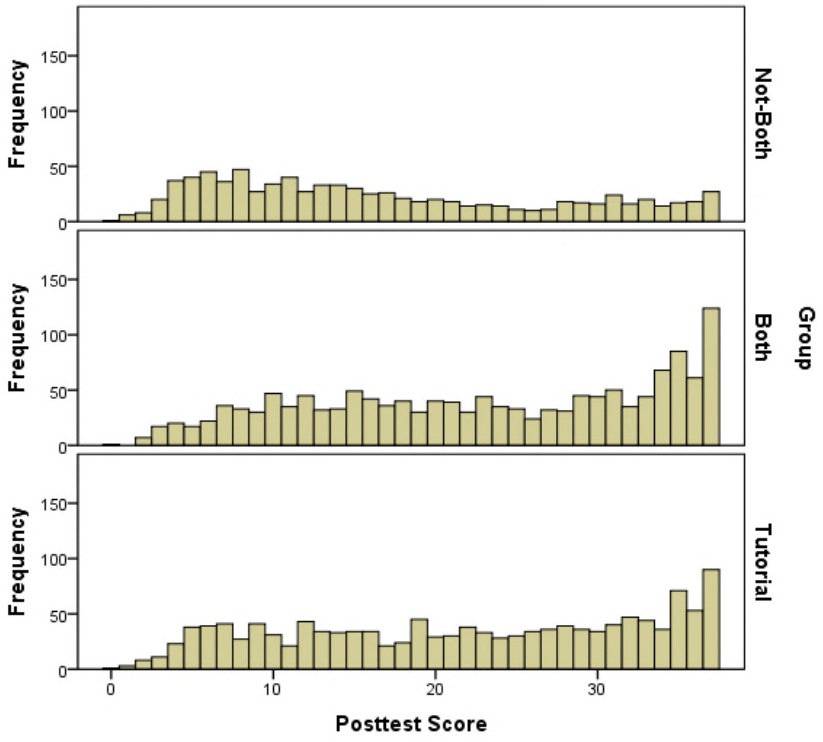


Figure 5-2. Distribution of total FMCE score at posttest by treatment group.

The narrative is confirmed by the plot of complete, matched pretest and posttest scores in Figure 5-3. The scatterplot seems to show two overlapping groups of students. The first group, about two-thirds of the sample, began the semester with very low scores and made gains that vary widely from 'no change' to 'hitting the ceiling.' The remaining third of students, those that started the semester with scores above 10, all made tremendous gains. These students overwhelmingly end the semester near the ceiling of the assessment. The implication is that college physics courses have an unpredictable effect on students with strong misconceptions, but a predictably positive effect on students who have shaken those misconceptions to some extent. It may be that students need at least two exposures to introductory physics to replace misconceptions with Newtonian thinking. This pattern is referred to as *staged learning* in this dissertation, and is similar to the learning progressions described in the science education literature (Wilson, 2009).

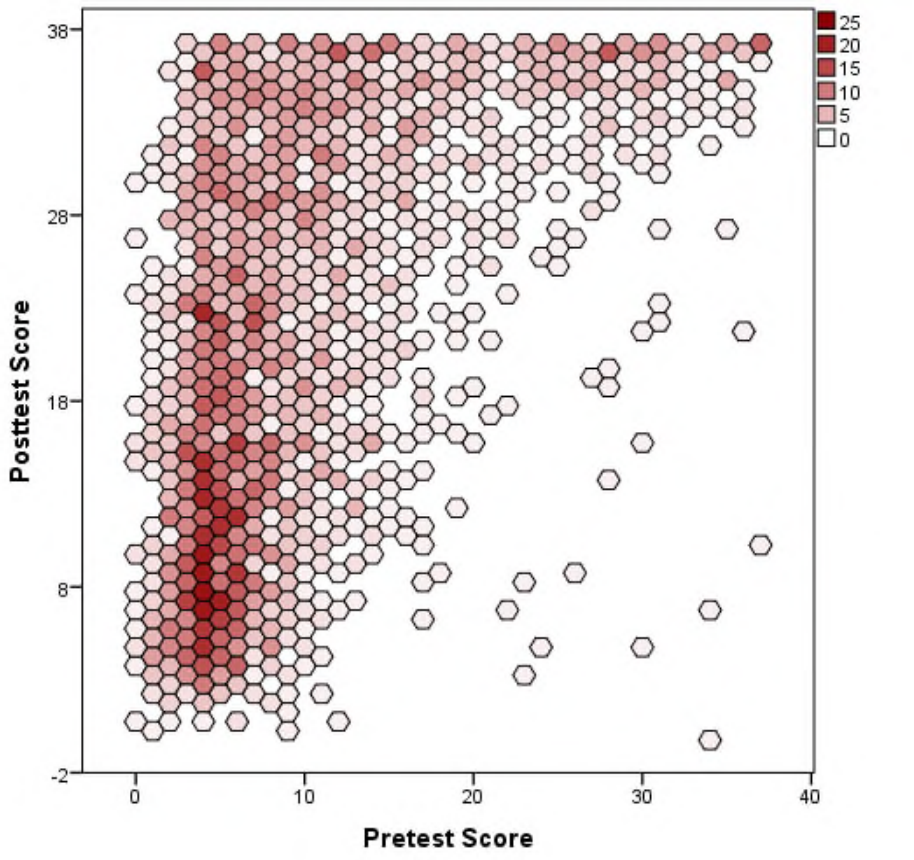


Figure 5-3. Weighted scatterplot of posttest FMCE scores against pretest FMCE scores. Darker shading indicates a larger number of students with that pair of pretest-posttest scores.

Factor analytic methods were applied to the FMCE data set to explore the dimensionality of the instrument in a conventional context. These analyses are largely tangential to the current study so they will be described here very briefly, while the full results are presented in Appendix B. FMCE scores from completed assessments have a Cronbach's α of .939 at pretest and .956 at posttest, with mean inter-item correlations of .286 and .344 respectively. Classical test theory and item response theory analyses showed that the items tend to be too difficult for pretest students and highly discriminating at both time points. These analyses also revealed that the Velocity Graphs testlet is too easy and does not discriminate well. Exploratory factor analysis yielded a six factor solution, with Force Graphs and Force Sleds loading on the same factor at

pretest and Force Graphs and Acceleration graphs loading on the same factor at posttest. The factors were fairly strongly correlated with one another, so in addition to confirmatory factor analysis, a bifactor model was fitted to the data. The bifactor model, which includes a latent variable for general ability and a set of latent variables for specific factors, fit the data significantly better than any other conventional model. Overall, the conventional modeling process indicates that the FMCE exists in the grey area between uni-dimensional and multi-dimensional scales.

Tests of Group Mean Differences

In a typical controlled trial study, pretest and posttest scores are entered into multiple regression models, with treatment conditions coded using dummy variables and pretest score as a covariate. With random assignment to groups, mean differences at the pretest would not be expected, nor would an interaction between pretest and group membership. However, strict random assignment was not used in this dataset, and as noted above, there were mean differences between the groups at the pretest. This by itself precludes strong causal inferences about treatment effects. Moreover, given the lack of random assignment, equal slope coefficients for the pretest across groups could not be assumed, so a test of interaction between treatment group and the pretest was included in the model.

The analysis described here uses the scoring template described in Chapter 4, includes a mean-centered version of the pretest score variable, and only include cases with complete FMCEs. The regression model is given by:

$$\text{Post} = B_0 + B_1(\text{Pre}) + B_2(\text{BOTH}) + B_3(\text{TUTORIAL}) + B_4(\text{Pre*BOTH}) + B_5(\text{Pre*TUTORIAL}) + \mu \quad (5\text{-}1)$$

where NOT BOTH is used as a reference group. The analysis was performed in SPSS Version 21. The model had an R^2 of .325, indicating that the independent variables describe approximately one-third of the variability in posttest FMCE scores. Table 5-2 describes the properties of the overall model while Table 5-3 enumerates the B coefficients of the multiple regression.

Table 5-2

ANOVA table for the multiple regression comparing scores across treatment groups

	Sum of Squares	Df	Mean Square	F	P
Regression	118394	5	23678	301	< .001
Residual	245192	3125	78		
Total	363587	3130			

Table 5-3

Coefficients of the multiple regression comparing scores across treatment groups

	Coefficient	Standard Error	T	P
(Constant)	9.333	.505	18.479	< .001
Pretest Score	0.907	.046	19.503	< .001
BOTH	5.125	.401	12.772	< .001
TUTORIAL	3.648	.427	8.537	< .001
Pre*BOTH	-0.157	.056	-2.773	.006
Pre*TUTORIAL	-0.251	.058	-4.306	< .001

The coefficients in Table 5-3 show that students in the BOTH group scored five points higher on average at posttest while the TUTORIAL students scored 3.6 points higher on average than students in the NOT BOTH group, after accounting for pretest score. The interaction terms

are statistically significantly different from zero and they are negative. This indicates that the impact of pretest score is greater in the NOT BOTH group than in either of the interventions, meaning that the successful students in the NOT BOTH group were likely to be those students who already had some understanding of the material at the beginning of the semester. These results indicate that the interventions may have been less dependent on students already entering the course with some conceptual knowledge.

While this analysis is substandard in terms of study validity and causal analysis because of lack of assignment, non-linearity, floor and ceiling effects, and implementation fidelity, the regression model shows that the groups are substantively different from one another. The BOTH and TUTORIAL groups had higher mean posttest scores after accounting for pretest scores. Taken at face value, the results would indicate that the interventions were more effective—but would provide no more detailed information. Conventional results can show that scores are better for some groups, but not what mental processes led to higher scores. The current study aims to answer the question of whether latent class modeling provides more detailed diagnostic data. These conventional tests are crucial because they show that the groups are meaningfully different from one another, a necessary step for testing the mLTA method.

The regression analysis above is a comparison of a composite score across groups. The latent class modeling, on the other hand, targets each testlet individually. To make a more direct comparison with conventional scoring, I have included histograms of testlet scores across groups and time points in Figure 5-4 through *Figure 5-9*. In each case, the students included were those that responded to at least one item on the testlet. Note that the NOT BOTH group was the smallest group and so the histograms have smaller peaks—comparisons can be made by looking at the proportional shape of each histogram. All figures show that students tended to score more

points at posttest than at pretest. All figures show that the students in each treatment group scored similarly at pretest (though statistically significantly different, as shown earlier).

The Force Sled and Reverse Direction histograms show the bimodal nature of performance on the FMCE at posttest. The majority of students either received no points or received all of the points. This hints at the 'they either get it or they don't' nature of conceptual testing and suggests that this kind of continuous scoring may not be appropriate. The floor and ceiling effects displayed in Figure 5-4 and **Error! Reference source not found.** may also drive the high Cronbach's α values mentioned in Chapter 4.

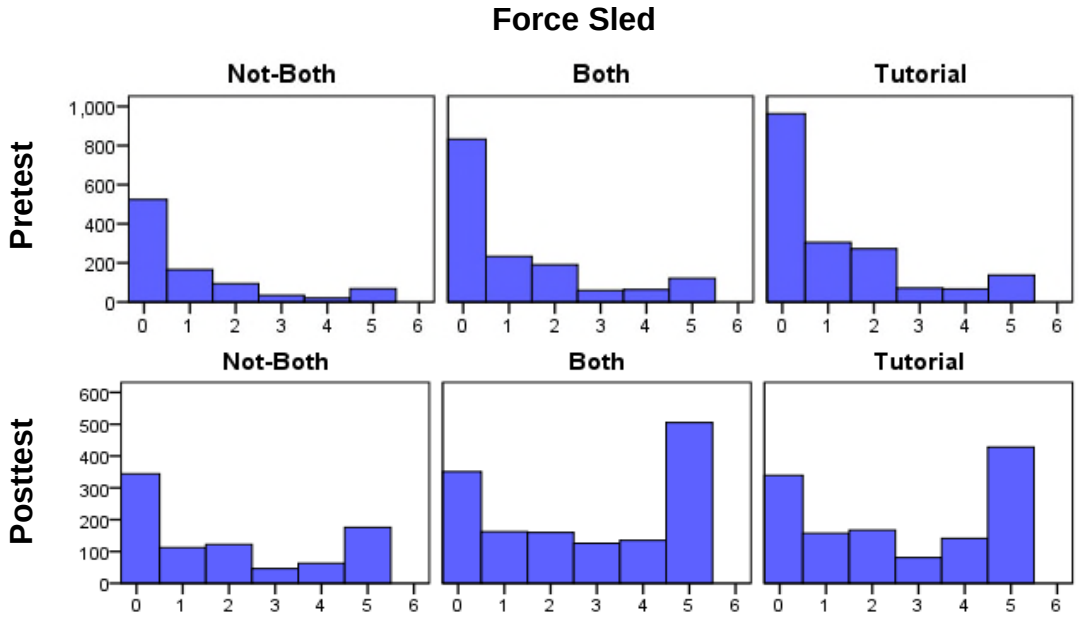


Figure 5-4. Histograms of pretest and posttest scores on the Force Sled testlet.

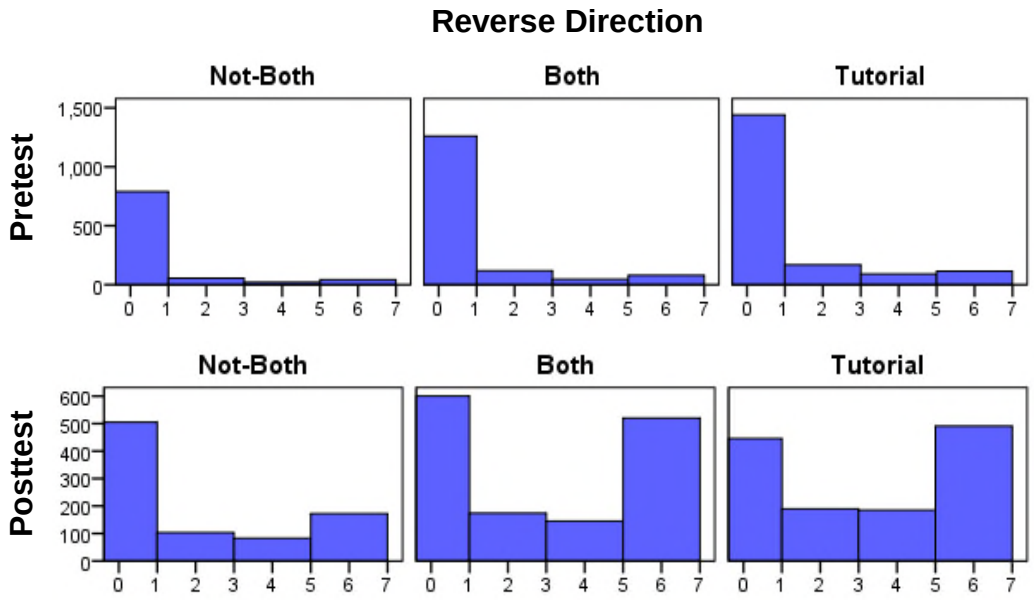


Figure 5-5. Histograms of pretest and posttest scores on the Force Sled testlet. Note that because of the scoring template described in Chapter 4, students may only score 0, 2, 4, or 6 points.

The Force Graphs and Acceleration Graphs histograms in Figure 5-6 and Figure 5-7 highlight the effect of the RTP/ILD and Tutorials intervention programs. In the posttest histograms, the students in the NOT BOTH group scored visibly lower than students in either of the intervention groups. The peaks of the NOT BOTH graphs are lower, but looking at the figures proportionally, they are visibly different shapes. This result further supports the assumption that the NOT BOTH group is substantively different from the other two, that the mLTA models should find differences across groups if they are as effective as regression or descriptive statistics.

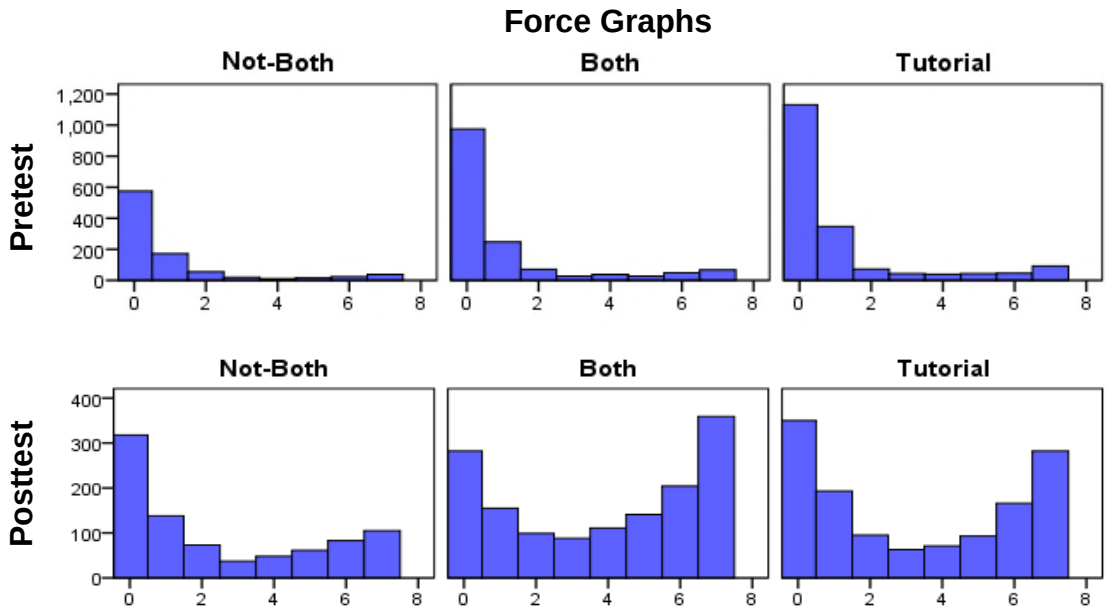


Figure 5-6. Histograms of pretest and posttest scores on the Force Graphs Testlet

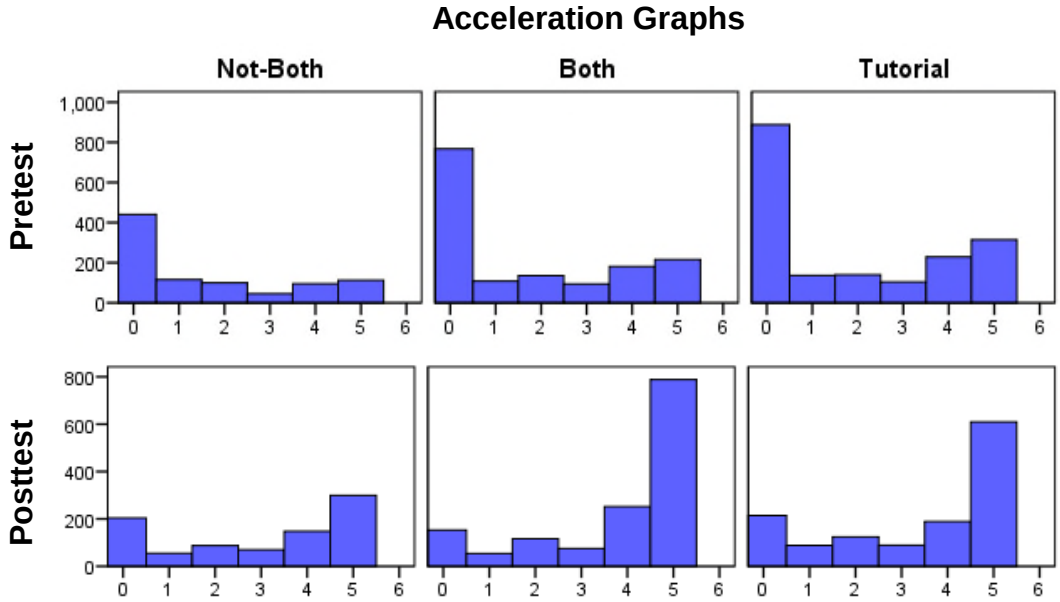


Figure 5-7. Histograms of pretest and posttest scores on the Acceleration Graphs testlet.

The Newton's Third Law and Energy testlet score histograms in Figure 5-8 and Figure 5-9 support the statements made above, but also highlight the odd scoring patterns caused by the structure of the FMCE. On the N3 testlet, particularly at posttest, students either scored zero,

four, or six points. The first three items on the testlet are linked by the scoring template to avoid false positives, so students only get credit for those three items if they get all three correct. Most students who get those three correct know the heuristic that 'the forces are equal in a collision' and so answer the fourth item correctly. Many students are able to answer the collision items correctly, scoring four points, but not the pushing questions, giving them just the four points. Some students have generalized the concept and score all six points. It is possible to get three or five points on the N3 testlet, but very few do because the items are highly related. This suggests, again, that a categorical scoring method is more appropriate. The Energy testlet is odd in that it seems somewhat continuous at pretest but has the same sort of jagged structure at posttest. The jagged structure is likely due to the fact that the testlet is comprised of two pairs of related items.

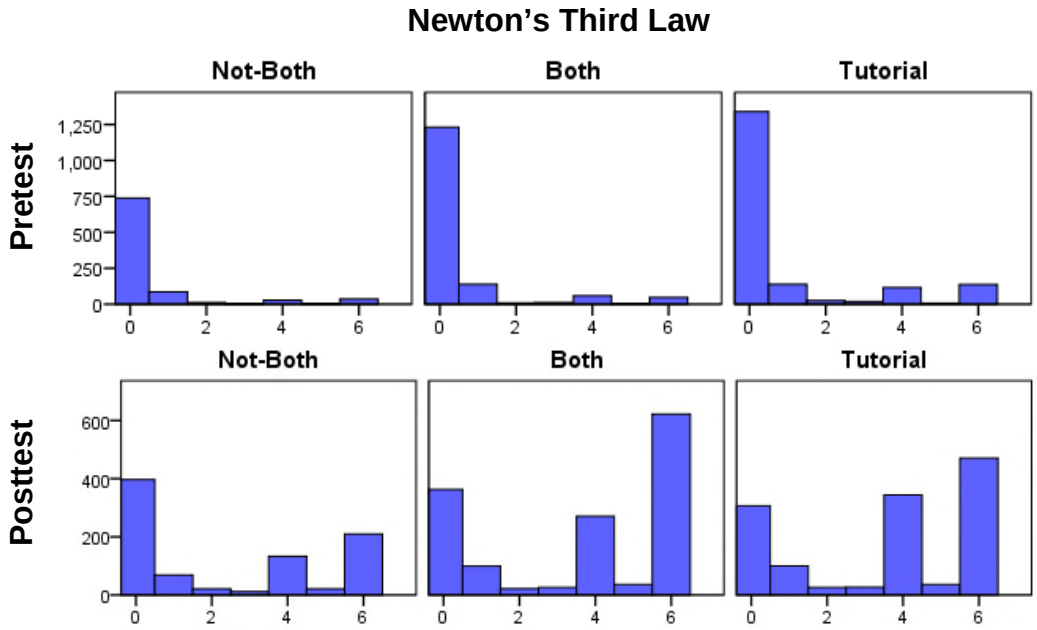


Figure 5-8. Histograms of pretest and posttest scores on the Newton's Third Law testlet.

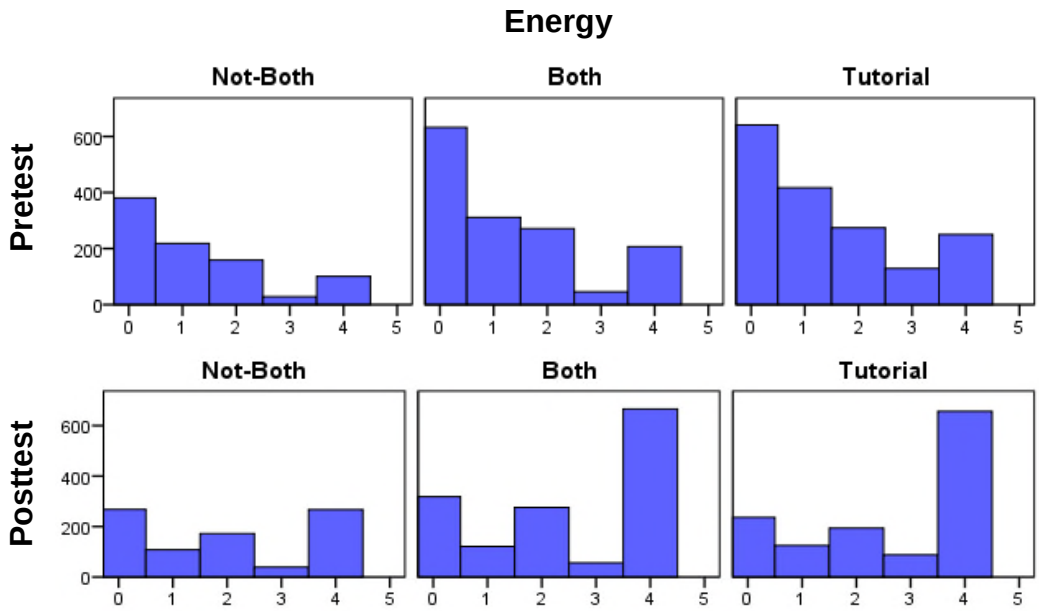


Figure 5-9. Histograms of pretest and posttest scores on the Energy testlet.

Latent Class Modeling

Force Sled

The Force Sled testlet, summarized in Table 5-4, asks students to select the force that matches the described motion of a sled across frictionless ice. In Newtonian physics, a net force on an object causes a change in the velocity of the object. A constant force to the right is then necessary to increase a velocity to the right or to decrease a velocity to the left. Students tend to confuse force with velocity, assuming a direct relationship. Some students express hybrid conceptions where they have internalized the directionality of forces and motion but still select the incorrect magnitudes.

Table 5-4

The Force Sled testlet of the FMCE, in abbreviated form.

Testlet Stem:	These items ask students to select a verbal description of a force that matches the motion of a sled across frictionless ice in one dimension.
Item	
1	Moving to the right, speeding up
2	Moving to the right, constant speed
3	Moving to the right, slowing down
4	Moving to the left, constant speed
7	Moving to the left, slowing down
Response	
A	Force to the right and increasing
B	Force to the right and constant
C	Force to the right and decreasing
D	No force
E	Force to the left and decreasing
F	Force to the left and constant
G	Force to the left and increasing
J	No response is correct

The top ten response patterns in the pretest and posttest data, shown in Table 5-5, include the correct response of BDFFB and the dominant misconception of ABCGE. The table also includes the hybrid ABFGB response, where students express the correct direction of the force but still have the force/speed confusion in terms of magnitude. The ABFGB students think that a constant force is necessary to slow an object down but an increasing force is necessary to speed it up. The response patterns ABEGC and ABGGA are similar hybrid responses, but believe that slowing an object down requires a decreasing or increasing force, respectively.

Table 5-5

Ten most common response patterns on the Force Sled testlet at pretest and posttest.

PRE	% Students	POST	% Students
BDFFB	0.08	BDFFB	0.31
ABCGE	0.32	ABCGE	0.16
ABFGB	0.08	ABFGB	0.07
ABCGB	0.04	ABCGB	0.03
ABGGA	0.03	ABGGA	0.03
ABCGA	0.02	ABCGA	0.01
BCFFB	0.01	BCFFB	0.02
ABCGF	0.02		
ABCGD	0.02		
ABEGC	0.02		
		ADFGB	0.02
		BDFGB	0.01
		ABFGA	0.01
Total	0.64		0.67

Table 5-6 shows a summary of the convergence, identification and model fit for all of the latent class models. For each model, the table includes the solution with the best BIC, a model with one fewer class and a model with one more latent class. The column labeled Rep. indicates whether the best fitting solution was found by at least two of the converged replications. The number of solutions that converged, out of twenty, is shown in the Conv. column. The log likelihood and BIC are measures of model fit and entropy indicates the degree of class separation. The error column uses the abbreviation NPD for a non-positive definite matrix error and is always followed by the condition number provided by Mplus, which indicates problems in model identification when smaller than 10^{-10}. The abbreviation IM indicates that some parameters were automatically fixed by Mplus to allow for an inversion of the information matrix, usually an indicator of extreme parameter values.

Table 5-6 shows that the pretest LCA models had identification issues, though the same issues were not found in the other models of the analysis. The entropy values in excess of 0.8 indicate an acceptable degree of class separation and discreteness in classification. The best multi-group latent transition analysis model produced a non-positive definite matrix error, though it was likely to be caused by extreme values rather than a problem with model identification, given the condition number in the 10^{-8} range.

A likelihood ratio test shows that the mLTA model fits the data significantly better than a constrained model where transition parameters were held constant across groups. In this case, the likelihood ratio test is 2(-33223 – -33307) = 168 and the difference in number of parameters is 317 – 257 = 40. The value 168, with 40 degrees of freedom, is statistically significant with a p-value less than .001. This result indicates that the transition parameters differ significantly across groups. However, the constrained model has a smaller BIC. It is unclear which of the two models fits the data better.

Table 5-6

Model estimation summary for the Force Sled testlet.

Model	Classes	Rep.	Conv.	LL	BIC	Entropy	Errors
Pre	5	Yes	17	-16030	33555	0.891	NPD e-12, IM
	6	No	13	-15846	33487	0.905	NPD e-14
	7	No	17	-15702	33499	0.894	NPD e-14
Post	3	Yes	20	-13499	27850	0.904	
	4	Yes	20	-13112	27364	0.905	
	5	No	13	-12983	27393	0.908	NPD, IM
Both	7	No	15	-29275	60802	0.894	NPD e-13, IM
	8	Yes	16	-29108	60789	0.905	IM
	9	No	15	-28978	60853	0.897	NPD e-11
LTA	7	Yes	19	-28276	59076	.800	IM, NPD e-14
	8	No	18	-28060	59011	.804	IM
	9	No	12	-27904	59138	.815	IM, NPD e-12
mLTA	5	Yes	20	-33522	69143	.836	IM
	6	Yes	16	-33223	69118	.846	NPD e-8, IM
	7	No	18	-32970	69235	.848	NPD e-16, IM
Constrained	6	No	17	-33307	68780	.843	IM

The classes identified in the LCA models are shown in Table 5-7 in terms of their modal responses. The classes include the correct BDFFB class, the dominant misconception ABCGE class, the hybrid conception ABFGB class, and several ‘other’ classes. Latent classes are labeled as ‘other’ if they lack identifying features, lack a coherent interpretation, or represent an amalgam of response patterns with different meanings. The quotation marks are a deliberate reminder that the classes are likely to be made up of random leftovers in the sorting process but should not be dismissed entirely. It is important to consider what meaning these ‘other’ classes might have.

One example of an 'other' class is AB*G*, which appears to be a coalition of all the observed response patterns that are similar to the common misconception but are not ABCGE or ABFGB. That particular class is defined more by what it is not. The measurement parameters, not shown here to conserve space, contain information on the probability of each response for each group rather than just naming which parameter is greatest. According to the full set of measurement parameters, Item 3 is *not* C or F and Item 7 is *not* E or B. Those parameters are fixed at Mplus' lower limit of -15. Similarly, the B**** class is made of the students who got the first answer correct but did not respond consistently with the correct answers to other items. The first might be called an 'other incorrect' class and the second an 'other correct' class. It is also common for LCAs to produce a class that is made up of the random leftovers of the data sample, shown in the combined LCA results in Table 5-7as *****.

Table 5-7

Classes identified by modal responses in pretest, posttest, and combined LCAs of the Force Sled testlet

Pre LCA	%	Post LCA	%	Combined	%
BDFFB	13	BDFFB	41	BDFFB	24
ABCGE	48	ABCGe	28	ABCGE	37
ABFGB	18	ABfGb	23	ABFGB	17
AB*G*	14			ABgGA	6
B****	5	Bd*f*	8	B*cfe	4
gFeAc	1			B**F*	4
				ABeG*	3
				*****	1

The combined data LCA results presented in Table 5-7 are generated from the combination of all pretest and posttest responses. In all the testlets, the combined analyses yielded better fit with a larger number of classes than either the pretest or posttest models, likely

because of the much larger amount of information available. It is possible that with a much larger and more diverse sample, some smaller classes reach a 'critical mass' such that the model fits better if it splits that class off from the others. In the case of the Force Sled data, the combined LCA produced the ABeG* class which likely includes the hybrid ABEGC students along with some other response patterns. Identifying the ABEGC students as separate allowed for the AB*G* group to become the other hybrid conception class ABgGA. Note that the sequential language, with terms like 'combining' and 'splitting,' is common among descriptions of latent class solutions—though it is inaccurate. The three models in Table 5-7 are independent solutions to the task of how to sort FMCE Force Sled results. The AB*G* class did not actually 'become' any other class.

The classes of the LTA and mLTA models are shown in Table 5-8. The LTA model arranges the 'other' classes somewhat differently from the combined LCA model, though the overall structure seems similar. It includes a AB*Gc class rather than a ABeG* class. Both versions of the class probably include the ABEGC students but combine them with different response patterns. As with all of the testlets, the mLTA model provided the best fit with fewer classes than the LTA model. One possible explanation for this is that the grouping variable provides information which the model can capitalize on, so fewer parameters and classes are necessary to describe the data. The mLTA solution seems very simple and clear. The five classes are correct, common misconception, hybrid conception, 'other correct,' 'other incorrect,' and 'other.'

Table 5-8

Classes identified by modal responses from LTA and mLTA of the Force Sled testlet

LTA	% Pre	% Post	mLTA	% Pre	% Post
BDFFB	10	40	BDFFB	12	41
ABCGE	46	22	ABCGE	48	23
ABfGB	20	18	ABFGB	19	18
B**F*	4	3	B**f*	5	7
*****	2	1	*****	2	1
			AB*G*	14	9
B*cfe	3	5			
AB*Gc	8	3			
ABgGA	8	7			

The transition probabilities of the LTA model are shown in Table 5-9, where each value is the probability of membership in a posttest class (column), given membership in a pretest class (row). In other words, each row shows the complete set of outcomes for students belonging to each of the eight classes at pretest. Note that the classes in Table 5-9 are presented in a random order (the order that they were presented in the Mplus output). The color coding loosely identifies columns and rows by classes that are correct, are common misconceptions, or are dual/hybrid conceptions.

Table 5-9

Probabilities of transitioning into posttest classes (column) given class membership at pretest (row) on the Force Sled testlet.

		1	2	3	4	5	6	7	8
		B**F*	B*cfe	AB*Gc	ABgGA	ABfGB	BDFFB	ABCGE	*****
1	B**F*	0.158	0.000	0.005	0.074	0.158	0.553	0.025	0.027
2	B*cfe	0.105	0.177	0.035	0.014	0.104	0.144	0.420	0.000
3	AB*Gc	0.093	0.027	0.083	0.041	0.255	0.354	0.147	0.000
4	ABgGA	0.025	0.025	0.042	0.235	0.193	0.394	0.086	0.000
5	ABfGB	0.021	0.012	0.005	0.069	0.331	0.520	0.042	0.000
6	BDFFB	0.000	0.000	0.007	0.008	0.015	0.957	0.000	0.013
7	ABCGE	0.024	0.077	0.030	0.071	0.145	0.250	0.397	0.006
8	*****	0.000	0.176	0.020	0.138	0.170	0.113	0.363	0.021

The values in Table 5-9 show that students beginning the semester in the correct class were very likely to stay in the correct class. Students beginning in the hybrid class had a 33% chance to stay in that class and a 52% chance to change to the correct class. Those students who began the semester with the dominant misconception were most likely to stay in the same class (40%), not very likely to move to the correct response class (25%), while a small proportion (17%) moved to the hybrid conception class.

This testlet shows some amount of staged learning, where students tend to follow a multistage path from incorrect to correct thinking. Staged learning predicts that students with the misconception will either stay, move to a hybrid conception, or move directly to the correct conception. Those with the hybrid conception will either stay or move to the correct answer class. Those in the correct answer group are likely to stay. The Force Sled testlet is an excellent example of this pattern.

The transition probabilities from the mLTA solution, one set for each treatment group, are presented in Table 5-10. The probabilities show that the BOTH and TUTORIAL

interventions were more effective than the NOT BOTH condition. For all pretest classes, the intervention groups were more likely to have students transition into the correct response class. Common misconception students in the intervention groups were less likely to stay in the misconception class. The BOTH and TUTORIAL groups appear similar overall, though the BOTH group appears to have been more successful at moving hybrid concept students to the correct class. The probabilities, in row 4 and column 3 of each sub-table, show that the BOTH group moved 64% of ABFBG students to the correct response rather than the 42-47% of the other two groups.

This result could be interpreted as a sign that that the RTP/ILD curriculum is more effective than Tutorials in Introductory Physics curriculum. The inference is difficult to justify, given the bias and sources of error previously described. Moreover, the difference is impossible to test statistically. It is not possible to test the difference between two parameters in two different groups because they do not share the same reference groups. Mplus offers a probability parameterization, which does not use reference groups and might allow for comparison of transition parameters. Unfortunately, the mLTA models would not converge with probability parameterization. As such, I am unable to perform post hoc statistical tests. However, this does serve as an example of the kind of result that might be useful for future researchers trying to compare various interventions.

Table 5-10

Probabilities of transitioning into posttest classes (column) given class membership at pretest (row) on the Force Sled testlet, for each treatment group.

NOT BOTH

		1 ABFGB	2 ABCGE	3 AB*G*	4 *****	5 BDFFB	6 B**f*
1	ABFGB	0.380	0.009	0.174	0.000	0.421	0.016
2	ABCGE	0.132	0.530	0.117	0.008	0.122	0.090
3	AB*G*	0.220	0.115	0.262	0.000	0.368	0.035
4	*****	0.089	0.532	0.061	0.050	0.083	0.184
5	BDFFB	0.043	0.000	0.027	0.027	0.903	0.000
6	B**f*	0.046	0.443	0.066	0.000	0.222	0.223

BOTH

		1 ABFGB	2 ABCGE	3 AB*G*	4 *****	5 BDFFB	6 B**f*
1	ABFGB	0.312	0.026	0.017	0.005	0.640	0.000
2	ABCGE	0.152	0.370	0.073	0.007	0.296	0.103
3	AB*G*	0.175	0.063	0.212	0.000	0.463	0.087
4	*****	0.330	0.269	0.184	0.000	0.074	0.143
5	BDFFB	0.014	0.000	0.015	0.000	0.971	0.000
6	B**f*	0.169	0.221	0.047	0.030	0.322	0.210

TUTORIAL

		1 ABFGB	2 ABCGE	3 AB*G*	4 *****	5 BDFFB	6 B**f*
1	ABFGB	0.330	0.062	0.078	0.000	0.478	0.055
2	ABCGE	0.171	0.364	0.068	0.012	0.308	0.077
3	AB*G*	0.234	0.122	0.174	0.008	0.455	0.007
4	*****	0.000	0.000	0.371	0.210	0.420	0.000
5	BDFFB	0.038	0.040	0.000	0.029	0.892	0.000
6	B**f*	0.127	0.093	0.042	0.033	0.536	0.169

Reverse Direction

The Reverse Direction testlet, summarized in Table 5-11, is the combination of three smaller testlets that all involve an object changing direction as it rises and falls. In all cases, the only force acting on the object is gravity. The correct response pattern is AAA AAA AAA responses. The dominant misconception is that the force on an object is directly related to the motion, so an object has an upward decreasing force as it rises, zero force at the top of the motion, and a downward increasing force as it falls. A student completely consistent with this conception will answer GDB GDB GDB.

Table 5-11

The Reverse Direction testlet of the FMCE, in abbreviated form.

Testlet Stem:	These items ask students to describe the net force on an object as it moves upward, stops, and comes back down. The first stem describes a car pushed up a ramp, the second two stems describe a coin toss.
Item	
8	The **force** on a **car** as it moves up a ramp
9	The **force** on a **car** at the top of its motion
10	The **force** on a **car** as it moves back down the ramp
11	The **force** on a **coin** as it goes up in the air
12	The **force** on a **coin** at the top of its motion
13	The **force** on a **coin** as it falls back down
27	The **acceleration** of a **coin** as it goes up in the air
28	The **acceleration** of a **coin** at the top of its motion
29	The **acceleration** of a **coin** as it falls back down
Response	
A	The force is down and constant
B	The force is down and increasing
C	The force is down and decreasing
D	The force is zero
E	The force is up and constant
F	The force is up and increasing
G	The force is up and decreasing
J	None is correct

The response patterns in Table 5-12 show that the Reverse Direction testlet is the most clear-cut example of a dual conception. Students that respond GDB AAA AAA or GDB GDB AAA clearly have the misconception and the correct conception stored in their brains but the activation and application of those conceptions is inconsistent. It is possible that the students who respond with GDB GDB AAA responded reflexively to the first two sets of items, giving the misconception answers, but were primed by the stimuli of the intervening items to give the correct answer for the final set of items. On the other hand, it is possible that the students responded to the context clues of each item stem and the differing answers were deliberate. Either way, the students demonstrated capacity for both ways of thinking, so these response patterns are indicators of a dual conception.

Table 5-12

Ten most common response patterns on the Reverse Direction testlet at pretest and posttest.

Pre	% Students	Post	% Students
AAAAAAAAA	0.05	AAAAAAAAA	0.30
GDBGDBGDB	0.26	GDBGDBGDB	0.08
GDBGDBAAA	0.02	GDBGDBAAA	0.03
GDBGDBCDF	0.02	GDBGDBCDF	0.01
GDBGDBADA	0.01	GDBGDBADA	0.01
FDBGDBGDB	0.02		
GDBGDBADE	0.01		
GDBGDBGDA	0.01		
GDBGDBEDA	0.01		
FDBFDBFDB	0.01		
		GDBAAAAAA	0.02
		AAAAAAADA	0.01
		ADAADAADA	0.01
		EDAAAAAAA	0.01
		ADAAAAAAA	0.01
Total	0.42		0.49

The top ten most frequent response patterns in Table 5-12 only account for 42% of pretest and 49% of posttest responses. There are a huge number of observed response patterns in the Reverse Direction testlet. Most response patterns account for less than one percent of the sample. This is, to some extent, a consequence of the large number of items and response options on each item. The contingency table for this testlet has a massive $8^9 = 1.3 \times 10^8$ cells. On one hand, latent class modeling is ideal for this testlet, which can group similar response patterns and present them in an interpretable format. On the other hand, the huge contingency table and number of items, responses, and classes interfered with model convergence and identification.

The modeling summary in Table 5-13 shows that almost all of the latent class models had non-positive definite matrix errors with very small condition numbers. The multi-group models would not converge at all. Because the convergence and identification problems may have been due to the huge number of parameters, as predicted by experts such as Collins and Lanza (2010), the RD models were run using just six items (8-13). The six item versions produced the same problematic results. It may be that there is not enough information to estimate the overall model, but also possible that there are latent classes with so few individuals that there is not enough information to estimate parameters for specific classes. It is also possible that the identification issues are due to violations of conditional independence (Berzofsky, Beimer, & Kalsbeek, 2014). The RD items may have more in common with the other items in their subset than the other RD items, beyond that which can be modeled by the LCA. The failure to model the Reverse Direction testlet may be used to inform future test design. The mLTA method may require testlets with smaller contingency tables, fewer latent classes, or less intra-testlet structure.

Table 5-13

Model estimation summary for the Reverse Direction testlet.

Model	Classes	Rep.	Conv.	LL	BIC	Entropy	Errors
Pre	7	No	19	-27406	58545	.886	
	8	No	18	-27123	58513	.896	IM 1p
	9	No	19	-26922	58646	.904	NPD e-10, IM
Post	6	Yes	20	-22921	48931	.915	NPD e-17, IM
	7	No	20	-22640	48885	.926	NPD e-18, IM
	8	No	20	-22428	48977	.921	NPD e-17
Both	10	No	20	-50490	106711	.905	NPD, e-18, IM
	11	No	19	-50201	106706	.909	NPD, 1-18, IM
	12	No	14	-49951	106781	.914	NPD, e-17, IM
LTA	9	No	20	-50860	107248	.817	IM
	10	No	20	-50450	107129	.820	NPD, e-15, IM
	11	No	5 (of 10)	-50261	107408	.825	NPD, e-15, IM
mLTA			0 0 0				

While all models had identification issues, the LCA and LTA models did converge on solutions. It is worth displaying and discussing the results, though they should be considered with some skepticism as the models are not fully identified. The modal responses of each class are shown in

Table 5-14 and demonstrate a wide variety of response patterns. All solutions included one class for the correct answer, one for the dominant misconception, and at least one for the dual conception described above. The pretest classes included a class of students who selected F (up and increasing) for objects as they rose, another class for students who selected A (down and constant) for objects as they fell, and a series of ‘other’ classes. The combined pretest and posttest LCA appears to be an appropriate combination of the two other solutions. The

combined data, however, also includes a new dual class where items 11-13 were answered correctly but the other mini-testlets were answered incorrectly.

Table 5-14

Classes identified by modal responses in pretest, posttest, and combined LCAs of the Reverse Direction testlet

Pre LCA	%	Post LCA	%	Combined	%
AAAAAAAAA	7	AAAAAAAAA	35	AAAAAAAAA	20
GDBGDBGDB	41	GDBGDBgDB	21	GDBGDBGDB	28
GDBGDBAdA	9	GDBGDBAaA	9	GDBGDBAaA	8
fDbfDb*DB	11			FDbfDB*DB	7
GDBGDBcDF	6			GDBGDBcDf	7
gDagDAgDa	9			gDagDAgDa	7
		*DbAAAAAA	9	gDbAAAAAA	6
		*d**D**d*	6	*d**d**d*	4
		G*BgAb*ab	6	G*bgaB*ab	4
		*DaadA*Da	12	*Da*DAAdA	6
				*D*AAA*Db	4
gD*aaABAB	7				
*Db*d*gDb	8				

The latent transition analysis classes in Table 5-15 agree with those of the combined LCA solution, though the best-fitting LTA had one less class. One interesting class to note is the fDb*Db*Db class, which appears to be defined less by what the responses are and more by what the responses are not. Members of that class all responded that there was no force on the object at the top of its motion but gave a variety of responses to the force on a rising object. This group appears to be differentiated by the fact that the first response for each stem was *not* G.

Table 5-15

Classes identified by modal responses from LTA of the Reverse Direction testlet

LTA	% Pre	% Post
AAAAAAAAA	6	35
GDBGDBGDB	41	12
GDBGDBAaA	10	9
fDb*Db*Db	12	4
GDBGDBcDf	9	5
gDagDAgDa	9	6
*D*AAAAAA	5	14
bB***	3	10
gJBgDb*Db	1	1
*D*AaA*Db	4	4

Table 5-16 contains the transition parameters from the Reverse Direction LTA and may be difficult to read as the table is split in half to fit the 10 x 10 matrix of transitions. The large number of transitions is difficult to interpret, especially because the probabilities spread across ten possible outcomes tend to be small. Students who began the semester in the dominant misconception class tended to either stay the same (19%), move to two of the dual conceptions (12-13%) or move to the correct conception (26%). Those students who began with the hybrid or dual conceptions were highly likely to move to the correct conception. Those with the correct conception were very highly likely to stay. The transition probabilities indicate some amount of staged learning, though it is impossible to make any strong claims, given the complex data structure and the identification issues.

Table 5-16

Probabilities of transitioning into posttest classes (column) given class membership at pretest (row) on the Reverse Direction testlet.

	1 GDBGDBGDB	2 **b**b***	3 gJBgDb*Db	4 *D*AaA*Db	5 GDBGDBcDf
GDBGDBGDB	0.190	0.115	0.005	0.055	0.067
bb***	0.055	0.307	0.013	0.058	0.000
gJBgDb*Db	0.000	0.110	0.000	0.072	0.066
*D*AaA*Db	0.018	0.059	0.008	0.029	0.007
GDBGDBcDf	0.173	0.134	0.004	0.046	0.114
fDb*Db*Db	0.165	0.150	0.000	0.028	0.061
GDBGDBAaA	0.012	0.028	0.000	0.021	0.010
AAAAAAAAA	0.000	0.022	0.000	0.000	0.005
*D*AAAAAA	0.007	0.049	0.000	0.000	0.011
gDagDAgDa	0.055	0.036	0.000	0.081	0.046

	6 fDb*Db*Db	7 GDBGDBAaA	8 AAAAAAAAA	9 *D*AAAAAA	10 gDagDAgDa
GDBGDBGDB	0.019	0.117	0.264	0.131	0.039
bb***	0.018	0.000	0.295	0.244	0.010
gJBgDb*Db	0.095	0.172	0.258	0.094	0.133
*D*AaA*Db	0.000	0.037	0.631	0.209	0.002
GDBGDBcDf	0.015	0.104	0.193	0.149	0.068
fDb*Db*Db	0.202	0.051	0.068	0.110	0.166
GDBGDBAaA	0.008	0.141	0.596	0.174	0.009
AAAAAAAAA	0.000	0.007	0.925	0.041	0.000
*D*AAAAAA	0.000	0.027	0.728	0.180	0.000
gDagDAgDa	0.005	0.065	0.342	0.221	0.149

Force Graphs

The Force Graphs testlet, summarized in Table 5-17, asks students to choose the graph that best describes a particular motion. The correct response set, EAEBBGE is selected by students who know that a force on an object causes a change in velocity. So any object speeding up to the right (or slowing down while moving to the left) has a net force to the right. If the

change in velocity is ‘steady’ then the force is a constant. The dominant misconception, where students believe that force and speed are directly related, leads students to choose ACBHDF*, where they would select an increasing force when the velocity is described as increasing.

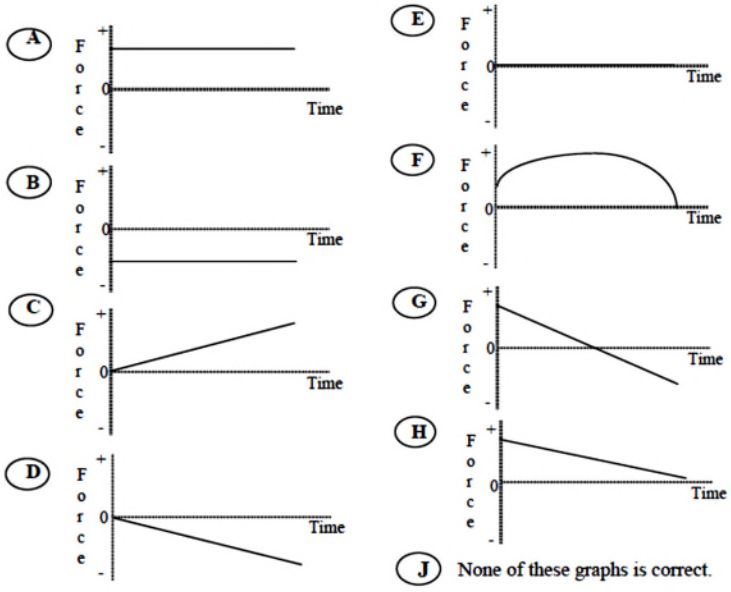
Table 5-17

The Acceleration Graphs testlet of the FMCE, in abbreviated form.

Testlet Stem:	These items ask students to consider a car moving in one dimension. Each item describes a motion and each response is a graph of force vs. time.

Item	
14	The car moves to the right, constant velocity
16	The car moves to the right, speeding up
17	The car moves to the left, constant velocity
18	The car moves to the right, slowing down
19	The car moves to the left, speeding up
20	The car moves to the right, speeds up then slows down
21	The car moves to the right, asks force after it is released

Responses



The most popular response patterns in Table 5-18 show that students have a wide range of responses to the last item in the testlet. Item 21 describes the toy car being pushed and then released and asks what graph describes the force on the car *after* it is released. This item was written to elicit the 'impetus' misconception described in Chapter 2. The responses in Table 5-18 do not implicate any particular hybrid or dual conception.

Table 5-18

Ten most common response patterns on the Acceleration Graphs testlet at pretest and posttest.

PRE		Post	
EAEBBGE	0.04	EAEBBGE	0.20
ACBHDFH	0.11	ACBHDFH	0.04
ACBHDFA	0.04	ACBHDFA	0.03
ACBHDFJ	0.03	ACBHDFJ	0.01
ACBHDFG	0.03	ACBHDFG	0.02
ACBHDFF	0.03	ACBHDFF	0.01
ACBHDFE	0.04	ACBHDFE	0.04
ACBGDFH	0.02		
ACBHDGF	0.01		
ACAHCFH	0.01		
		EAEBBFE	0.04
		EAEBBJE	0.03
		EABBBGE	0.01
	0.36		0.43

Table 5-19 gives a summary of model convergence, identification and fit. The values in the Conv. column are out of 12 replications, rather than 20, because the preliminary analyses were satisfactory. Other testlets were updated after the decision to have Mplus attempt to converge 20 replications. The combined LCA, using both pretest and posttest data, had a non-positive definite matrix error with a condition number in the 10^{-11} range, indicating some issue with model identification. The same errors did not occur with the LTA or mLTA models. A

likelihood ratio test shows that the mLTA model fit better than its constrained counterpart (206, 50 df, $p < .001$). However, the constrained model has a lower BIC. It is unclear which model fits the data better.

Table 5-19

Model estimation summary for the Acceleration Graphs testlet.

Model	Classes	Rep.	Conv.	LL	BIC	Entropy	Errors
			(Of 12)				
Pre	4	Yes	12	-29002	59900	.858	NPD e-10
	5	Yes	12	-28733	59838	.876	IM
	6	Yes	11	-28521	59890	.877	NPD e-15
Post	3	Yes	12	-23793	48979	.908	
	4	Yes	12	-23444	48748	.890	
	5	Yes	10	-23229	48786	.897	IM
Both	6	Yes	12	-52876	108810	.885	
	7	Yes	11	-52560	108689	.894	NPD e-11
	8	Yes	8	-52312	108704	.891	
			(Of 20)				
LTA	6	Yes	20	-51690	106506	.792	IM
	7	Yes	18	-51363	106435	.800	IM
	8	Yes	18	-51072	106450	.767	NPD e-10, IM
mLTA	5	Yes	18	-56853	116689	.834	NPD e-11, IM
	6	Yes	20	-56375	116483	.842	IM
	7	Yes	20	-56009	116552	.848	NPD e-12, IM
Constrained	6	Yes	20	-56478	116185	.841	IM

The classes in Table 5-20, labeled by their modal responses, confirm that there is no prominent dual or hybrid conception expressed through the Force Graphs testlet. Each model includes a correct answer class, a dominant misconception class, and one prominent class that

resembles the common misconception but with a variety of answers for items 17-19. There is a small group of students who are characterized mainly by selecting C for item 13, which has an increasing force graph for a constant velocity to the right. The students who selected C may have been thinking of a position vs. time graph, which is similar to the misconception except even more incorrect than assuming a velocity vs. time graph. The combined data appears to be an appropriate combination of the classes observed in the pretest and posttest data. It seems that the Cc***F* class was split into the C****F* class and the ****** class when the data was combined, though there may have been other 'shuffling' that is not apparent from the modal responses.

Table 5-20

Classes identified by modal responses in pretest, posttest, and combined LCAs of the Acceleration Graphs testlet

Pre LCA	%	Post LCA	%	Combined	%
EAEBBGE	10	EAeBBGE	13	EAEBBGE	23
ACBHDF*	52	ACBHDF*	53	ACBHDF*	40
AC***F*	22	AC***F*	22	AC***F*	18
Cc***F*	11	cc***f*	11	C****F*	5
gA**BgE	4			hA**BgE	3
				Eae**f*	7
				*******	3

The LTA classes in Table 5-21 closely resemble those of the combined LCA classes shown above. The mLTA model presents the same classes but appears to combine the 'other' group, characterized as *******, into the two classes beginning with C and H. The mLTA solution includes Eae**f*, students who tended to answer the first items correctly, and HA**BgE, students who answered the last items correctly. These class labels do not give a

coherent view of what the students might be thinking so these could be labelled as 'half-correct' instead of 'hybrid.' The HA**BgE class in the mLTA solution is strongly defined by their answer of H to item 14 and B to item 19. This particular response is unusual because items 14 and 19 ask about identical forces in opposite directions, yet the students answer them differently. It is likely that the latent class models identified these students specifically because they were so unique.

Table 5-21

Classes identified by modal responses from LTA and mLTA of the Newton 3 testlet

LTA	% Pre	% Post	mLTA	% Pre	% Post
EAEBBGE	8	36	EAEBBGE	8	36
ACBHDF*	52	29	ACBHDF*	52	29
AC***F*	25	7	AC***F*	24	8
Eae**f*	3	16	Eae**f*	3	17
C****F*	7	5	c****f*	9	8
hA**Bg*	3	3	hA**BgE	3	3
*******	3	4			

The transition probabilities in Table 5-22 show some fairly predictable results. Those students with the common misconceptions were likely to stay, move to the correct answer class, or move to the Eae**f* half-correct class. Those that began with Eae**f* were most likely to move to the correct class, sometimes staying in the same class, and sometimes moving to the common misconception class. This class appears to act as a hybrid class, though it lacks any specific identity to interpret. The students that began in the correct class were highly likely to stay in that class. The quirky students who initially answered Item 14 with C were likely to change their responses but not particularly likely to move to the correct answer class.

Table 5-22

Probabilities of transitioning into posttest classes (column) given class membership at pretest (row) on the Newton 3 testlet.

		1	2	3	4	5	6	7
		hA**Bg*	C****F*	ACBHDF*	Eae**f*	EAEBBGE	*******	AC***F*
1	hA**Bg*	0.102	0.000	0.055	0.061	0.783	0.000	0.000
2	C****F*	0.028	0.190	0.234	0.208	0.132	0.055	0.152
3	ACBHDF*	0.038	0.030	0.333	0.165	0.351	0.040	0.043
4	Eae**f*	0.015	0.000	0.164	0.194	0.567	0.048	0.013
5	EAEBBGE	0.019	0.000	0.010	0.009	0.938	0.023	0.000
6	*******	0.070	0.155	0.215	0.181	0.120	0.216	0.042
7	AC***F*	0.009	0.069	0.358	0.173	0.195	0.034	0.163

The transition probabilities for each treatment group are presented in Table 5-23 and show that the NOT BOTH condition, which approximates a control group, was less effective than the BOTH and TUTORIAL groups. The probabilities of transitioning into the dominant misconception were smaller and the probabilities of transitioning into the correct answer class were greater, as shown in columns 3 and 4. The BOTH group has more favorable transitions than the TUTORIAL group, though it is unclear whether the magnitudes of the transitions indicate a statistically significantly larger impact of the treatment. In particular, the BOTH group, taught by the RTP/ILD curriculum, was very successful in moving the half-correct Eae**f* class students to the correct response class. This effect can be seen in column 3, row 5 of each sub-table.

Table 5-23

Probabilities of transitioning into posttest classes (column) given class membership at pretest (row) on the Newton 3 testlet, for each treatment group.

NOT BOTH

		1 c****f*	2 hA**Bge	3 EAEBBGE	4 ACBHDF*	5 Eae**f*	6 AC***F*
1	c****f*	0.296	0.038	0.119	0.239	0.113	0.195
2	hA**Bge	0.000	0.081	0.764	0.039	0.116	0.000
3	EAEBBGE	0.026	0.031	0.899	0.017	0.027	0.000
4	ACBHDF*	0.054	0.036	0.224	0.483	0.163	0.041
5	Eae**f*	0.000	0.034	0.418	0.000	0.518	0.029
6	AC***F*	0.104	0.006	0.094	0.429	0.136	0.230

BOTH

		1 c****f*	2 hA**Bge	3 EAEBBGE	4 ACBHDF*	5 Eae**f*	6 AC***F*
1	c****f*	0.172	0.000	0.161	0.170	0.368	0.127
2	hA**Bge	0.000	0.088	0.853	0.017	0.043	0.000
3	EAEBBGE	0.000	0.011	0.967	0.000	0.021	0.000
4	ACBHDF*	0.032	0.043	0.435	0.250	0.210	0.031
5	Eae**f*	0.000	0.020	0.870	0.046	0.000	0.064
6	AC***F*	0.093	0.010	0.252	0.326	0.210	0.103

TUTORIAL

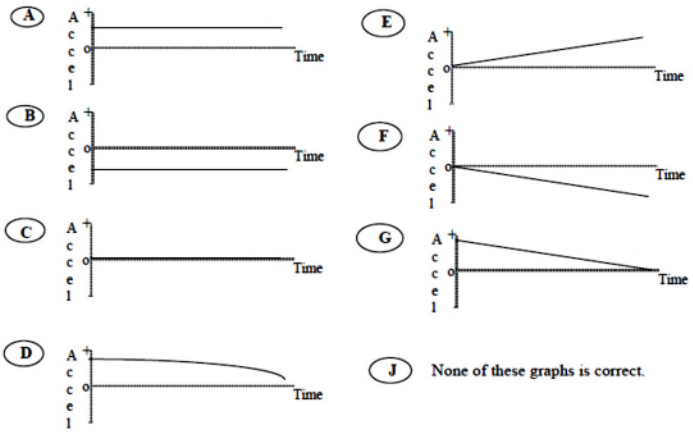
		1 c****f*	2 hA**Bge	3 EAEBBGE	4 ACBHDF*	5 Eae**f*	6 AC***F*
1	c****f*	0.383	0.051	0.090	0.327	0.150	0.000
2	hA**Bge	0.000	0.132	0.649	0.185	0.034	0.000
3	EAEBBGE	0.013	0.017	0.907	0.033	0.029	0.000
4	ACBHDF*	0.075	0.029	0.349	0.322	0.145	0.080
5	Eae**f*	0.000	0.012	0.453	0.229	0.273	0.033
6	AC***F*	0.078	0.012	0.218	0.326	0.210	0.115

Acceleration Graphs

The Acceleration Graphs testlet, summarized in Table 5-24, asks students to pick a graph that matches the described motion of a toy car. Since all of the velocities increase or decrease at a steady rate, all the correct responses use the A, B, and C options with constant acceleration over time. Any increasing velocity to the right or decreasing velocity to the left results from a constant acceleration to the right because acceleration is a rate of change in velocity. Students commonly mistake acceleration for speed, choosing graphs with a positive slope when the speed is increasing.

Table 5-24

The Acceleration Graphs testlet of the FMCE, in abbreviated form.

Testlet Stem:	These items ask students to consider a car moving in one dimension. Each item describes a motion and each response is a graph of acceleration vs. time.
Item	
22	The car moves to the **right**, **speeding** up
23	The car moves to the **right**, **slowing** down
24	The car moves to the **left**, **constant** velocity
25	The car moves to the **left**, **speeding** up
26	The car moves to the **right**, **constant** velocity
Responses	

The correct response pattern, shown in Table 5-25, is ABCBC, while the common misconception is EGBFA. The hybrid response pattern is EGCFC, where the students have apparently learned the heuristic that 'constant velocity means zero acceleration' but still choose the incorrect responses when the object is speeding up or slowing down. Table 5-25 also reveals that a relatively small number of response patterns account for a large proportion of students, 57% of pretest and 74% of posttest students.

Table 5-25

Ten most common response patterns on the Acceleration Graphs testlet at pretest and posttest.

PRE	% Students	POST	% Students
ABCBC	0.14	ABCBC	0.47
EGBFA	0.26	EGBFA	0.09
ABCAC	0.04	ABCAC	0.05
AGCBC	0.02	AGCBC	0.03
ABCFC	0.02	ABCFC	0.04
EGCFC	0.02	EGCFC	0.02
EGBJA	0.02		
EDBFA	0.02		
EGAEA	0.02		
EFBFA	0.01		
		AGBFC	0.01
		AGCFC	0.01
		AACBC	0.01
		EGBFC	0.01
	0.57		0.74

The modeling summary in Table 5-26 shows that the solutions have a high degree of class separation as expressed by the entropy. This may be due to the relatively small number of response patterns. The LTA model was probably not identified, given the non-positive definite matrix error and the small condition number in the 10^{-12} range. The solutions with one greater

and one fewer latent classes did not have the same identification warning. This presents an interesting dilemma in terms of model selection. It may be that the six-class solution is superior in that it is fully identified, though it has a greater BIC. The class structure of the two solutions offers little information as they appear very similar on the surface. Fortunately, the primary focus of the current study is on the mLTA results, so it is not necessary to deliberate between two awkward options. The mLTA model fits statistically significantly better than the constrained model, as determined by a likelihood ratio test (178, 56 df, $p < .001$). However, because the constrained model had a smaller BIC, it is unclear which model fits better.

Table 5-26

Model estimation summary for the Acceleration Graphs testlet.

Model	Classes	Rep.	Conv.	LL	BIC	Entropy	Errors
Pre	5	Yes	20	-18441	38377	.900	IM
	6	Yes	14	-18261	38317	.900	
	7	Yes	13	-18123	38342	.917	NPD e-11, IM
Post	3	Yes	20	-12552	25980	.907	
	4	Yes	20	-12278	25728	.917	
	5	Yes	14	-12146	25759	.928	IM
Both	6	Yes	13	-31279	64487	.910	NPD e-12
	7	Yes	9	-31071	64394	.912	
	8	Yes	8	-30951	64476	.917	
LTA	6	Yes	20	-30311	62687	.824	
	7	Yes	16/30	-30092	62653	.832	NPD e-12, IM
	8	Yes	18	-29939	62769	.826	IM
mLTA	4	Yes	20	-35886	73424	.842	IM
	5	Yes	20	-35268	72769	.856	IM
	6	Yes	20	-35143	72853	.865	IM
Constrained	5	Yes	20	-35357	72475	.859	IM

The modal responses of the classes are given in 5-27, where the three main classes in each model (correct, common misconception, and hybrid) account for 80% of all students. The hybrid class appears to be a coalition of responses, given the lower probabilities of response by individuals in the class. The lower case letters indicate a 0.5 to 0.7 probability of response. The egcfC class appears to be defined mostly by the very high probability of answering C to item 26. The 'other' groups in Table 5-27 appear to be defined by E or A for the first item and then a string of miscellaneous responses. It seems likely that they are defined more by *not* being ABCBC or EGBFA than by any particular response pattern. The combined model seems to be an appropriate combination of the pretest and posttest solutions, though the addition of extra information appears to allow the A**** to split into ***** and Ab*b*.

Table 5-27

Classes identified by modal responses in pretest, posttest, and combined LCAs of the Acceleration Graphs testlet

Pre LCA	%	Post LCA	%	Combined	%
ABCBC	28	ABCBC	64	ABCBC	44
EGBFA	36	EGBFA	14	EGBFA	25
egcfC	9	eGCFC	12	egcfC	10
Eg**A	10			Eg**A	7
E***a	8			E***a	5
A****	8	A****	9		
				Ab*b*	6
				*****	3

The LTA class profiles in Table 5-28 match exactly with those of the combined LCA analysis. The best fitting multi-group model has two fewer classes than the LTA model. As expected, the final model classes include the correct answer, the dominant misconception, the

hybrid conception, an ‘other correct,’ and an ‘other incorrect’ class. These match closely with the classes anticipated from the testlet’s most popular response patterns.

Table 5-28

Classes identified by modal responses from LTA and mLTA of the Acceleration Graphs testlet

LTA	% Pre	% Post	mLTA	% Pre	% Post
ABCBC	27	64	ABCBC	27	63
EGBFA	36	12	EGBFA	36	12
egcfC	10	11	egcfC	9	11
Eg**a	11	1	Eg**A	19	5
Ab*b*	6	3	A****	9	9
*****	3	5			
E***a	8	2			

The LTA transition probabilities in Table 5-29 show that the testlet has some amount of staged learning, though the results must be considered with some skepticism as the model was not fully identified. The students who began the semester in the correct class were very likely to stay in the correct class. Those that began in the hybrid class were very likely to move to the correct response. The students with the dominant misconception had a 57% chance of moving to the correct class, a 14% chance of moving to the hybrid class, and a 17% chance of staying in the same class. Students in the ‘other incorrect’ classes (E***a and Eg**A) had a reasonable chance of moving to the correct class (35-40%) but also a reasonable chance of moving to the dominant misconception (~20%). Very few students who began with the dominant misconception moved into the ‘other incorrect’ groups.

Table 5-29

Probabilities of transitioning into posttest classes (column) given class membership at pretest (row) on the Acceleration Graphs testlet.

	1 EGBFA	2 egcfC	3 E***a	4 Eg**A	5 ABCBC	6 Ab*b*	7 *****
EGBFA	0.174	0.138	0.023	0.013	0.572	0.044	0.035
egcfC	0.071	0.138	0.000	0.009	0.715	0.048	0.019
E***a	0.218	0.147	0.123	0.034	0.354	0.044	0.080
Eg**A	0.221	0.174	0.017	0.086	0.417	0.000	0.085
ABCBC	0.016	0.042	0.004	0.001	0.883	0.016	0.038
Ab*b*	0.052	0.051	0.026	0.000	0.769	0.069	0.034
*****	0.121	0.131	0.047	0.079	0.431	0.063	0.129

The transition parameters in Table 5-30, from the five class mLTA model, show the same pattern of staged learning. In all treatment groups, the Eg**A group was likely to transition to the dominant misconception class, which was likely to transition to the hybrid class, which was very likely to transition to the correct response class. The BOTH and TUTORIAL groups had greater probabilities of transitioning ‘up’ the hierarchy and smaller probabilities of staying in the same incorrect class. This indicates that those two interventions were more effective than the control condition, with the usual caveats about the validity of the comparison. It appears, particularly in examination of column 5 in each sub-table, that the BOTH group had more favorable transitions than the TUTORIAL group.

Table 5-30

Probabilities of transitioning into posttest classes (column) given class membership at pretest (row) on the Acceleration Graphs testlet, for each treatment group.

NOT BOTH

		1 egcfC	2 Eg***	3 A****	4 EGBFA	5 ABCBC
1	egcfC	0.176	0.000	0.058	0.126	0.640
2	Eg***	0.158	0.154	0.131	0.315	0.242
3	A****	0.137	0.086	0.132	0.096	0.549
4	EGBFA	0.148	0.051	0.104	0.275	0.422
5	ABCBC	0.048	0.000	0.037	0.020	0.895

BOTH

		1 egcfC	2 Eg***	3 A****	4 EGBFA	5 ABCBC
1	egcfC	0.117	0.000	0.019	0.011	0.853
2	Eg***	0.173	0.116	0.082	0.150	0.479
3	A****	0.072	0.012	0.116	0.030	0.769
4	EGBFA	0.102	0.015	0.062	0.122	0.695
5	ABCBC	0.030	0.000	0.033	0.005	0.931

TUTORIAL

		1 egcfC	2 Eg***	3 A****	4 EGBFA	5 ABCBC
1	egcfC	0.099	0.041	0.178	0.073	0.609
2	Eg***	0.116	0.144	0.152	0.241	0.347
3	A****	0.040	0.090	0.137	0.070	0.663
4	EGBFA	0.160	0.036	0.096	0.174	0.534
5	ABCBC	0.046	0.016	0.088	0.024	0.825

Newton's Third Law

The testlet that targets Newton's Third Law, summarized in 5-31, uses a hypothetical car and truck to explore student misconceptions. Forces between two objects are always equal, though it does not seem so in the real world because some objects 'hit harder' than others. In reality, the harder hitting object has more momentum and so has more influence in a collision or pushing situation. Objects with different momenta interact during the time when they are in contact by exerting equal forces on each other. Students tend to believe that the larger or faster moving object exerts a greater force during a collision, or sometimes believe that one object is the 'actor' and is the only one that can exert a force. When interpreting the class structure tables later in this section, note that the responses for items 36 and 38 are rearranged such that A is the correct answer. This is NOT reflected in 5-31, for simplicity of presentation.

Table 5-31

The Newton 3 testlet of the FMCE, in abbreviated form.

Testlet Stem:	The first four questions of the testlet use a stem where a car and truck have a head on collision. The last two questions ask about a situation in which the truck has broken down and the car is giving a helpful push.
Item	Asks students to compare the forces between a car and truck…
30	…if the two collide while moving the same speed
31	…if the two collide with the car moving much faster
32	…if the truck is standing still when the car collides with it
34	…if the truck is standing still and has the same mass as the car
36	…if the car is pushing the broken down truck, accelerating
38	…if the car is pushing the broken down truck, decelerating
Response	
A	The truck exerts a greater force
B	The car exerts a greater force
C	Neither exerts a force
D	The truck exerts a force but not the car
E	The truck and car exert the same forces
F	Not enough information to choose a response
J	No response is correct

The most popular response patterns to this testlet, displayed in Table 5-32, begin with the correct response set EEEEAA where all the responses indicate equal forces. The AFBBCB pattern represents the most common misconception where the larger or faster object exerts a greater force. Item 31 asks about a situation where the smaller object is moving faster, forcing students to apply both heuristics at the same time. The F in the most common misconception indicates that students believe the question is impossible to answer without numerical values. The ABBBCB response indicates that the student puts more importance on the speed of an object in a collision while AEBBCB assumes that the increased speed of the car compensates for its smaller size. The EEEECB response is correct for the collision items but incorrect for the pushing items. This is an interesting dual conception where the students have internalized the

correct concept for some contexts but not for others. The collision items are typical for introductory physics students but the pushing items are not typical, so it appears that some students revert to naïve intuitions in unfamiliar contexts.

Table 5-32

Ten most common response patterns on the Newton 3 testlet at pretest and posttest.

PRE		POST	
EEEEAA	0.03	EEEEAA	0.36
AFBBCB	0.22	AFBBCB	0.09
EEEECB	0.03	EEEECB	0.14
AFFBCB	0.07	AFFBCB	0.03
AEBBCB	0.07	AEBBCB	0.04
ABBBCB	0.05	ABBBCB	0.01
EBBBCB	0.02	EBBBCB	0.01
AFJJCB	0.02		
AFJBCB	0.02		
AFBBCC	0.01		
		EEEECA	0.01
		AFEBCB	0.01
		AFEECB	0.01
Total	0.54		0.71

The modeling summary in Table 5-33 shows that the combined data LCA, LTA, and mLTA models suffered from identification issues, with non-positive definite matrix errors and condition numbers in the 10^{-13} to 10^{-14} range. In some cases, the identification issues may stem from a single class, while in other cases Mplus error messages specified a single problematic transition parameter. It is difficult to know how much of the model is impacted by these localized problems—or if the localized problems are indicative of model-wide issues. I will present the results of these models, though they should be considered with some amount of skepticism. I will note here, and discuss further in Chapter 6, that this model has the same

violation of conditional independence as the Reverse Direction testlet. The mLTA model with seven classes fit statistically significantly better than its constrained counterpart, according to the likelihood ratio test (304, 84 df, $p < .001$). However, the lower BIC of the constrained model makes it unclear which model fits the data better.

Table 5-33

Model estimation summary for the Newton 3 testlet.

Model	Classes	Rep.	Conv.	LL	BIC	Entropy	Errors
Pre	6	Yes	20	-19373	40490	.850	IM
	7	Yes	16	-19221	40477	.865	
	8	Yes	10	-19103	40534	.853	NPD e-14
Post	4	Yes	18	-13631	28401	.951	
	5	Yes	18	-13484	28394	.903	
	6	Yes	18	-13358	28428	.913	IM
Both	7	Yes	19	-34014	70215	.898	
	8	Yes	16	-33851	70203	.887	NPD e-14
	9	Yes	19	-33709	70232	.894	
LTA	6	No	9	-32856	67725	.792	
	7	Yes	14	-32527	67464	.785	IM
	8	Yes	15	-32350	67523	.780	NPD e-12, IM
mLTA	6	Yes	18	-37459	77539	.845	IM
	7	Yes	17	-37108	77452	.834	NPD e-13, IM
	8	Yes	15	-36897	77695	.824	NPD e-14, IM
Constrained	7	Yes	17	-37263	77053	.833	IM

The classes of the LCA models, shown in Table 5-34, include the three most prevalent patterns described above. The ‘other’ classes of the pretest and combined data models include

AfJjCB, a class largely defined by the J response to item 32. The J response option, that 'none of these descriptions is correct,' may be distinct from option F, that 'there is not enough information to answer,' in the minds of the students or may be seen as equivalent. It is difficult to infer what these students were thinking. Another interesting class is aBBBCB in the pretest solution, which does not appear in the combined data model, though it accounts for about 13% of all pretest students. It seems likely that those students were split in the combined model into the A***CB, AfBBCB, and eBbCB classes. With so many similar response patterns, it is possible that students with different ideas are grouped inappropriately within the LCA model, though all of the students in these classes appear to have versions of the misconception. The entropy values for these models, as seen above, are strong in the LCA models, so the sorting of individuals is fairly distinct though it may be substantively subtle.

Table 5-34

Classes identified by modal responses in pretest, posttest, and combined LCAs of the Newton 3 testlet

Pretest	%	Posttest	%	Combined	%
EEEEAA	6	EEEEAA	39	EEEEAA	21
AFBBCB	53	AfBBCB	19	AfBBCB	32
EEEECB	5	EEEECB	23	EEEECB	14
Afe*CB	8	A**bCB	15	A***CB	7
AFJjCB	7			AfJjCB	5
A**b**	7			A*****	5
aBBBCB	13				
		******	4		
				AFFBCB	9
				eBbCB	7

The LTA and mLTA classes are shown in Table 5-35 and appear to sort the students differently. The three most important classes appear as distinct groups but the remaining 35 percent of the students were sorted differently across the different models. The AF**CB and AbbBCB LTA classes appear to be rearranged in the mLTA solution among the A*BBCB and AfeeCB classes. In that shuffle, approximately 9% of the pretest sample was moved from the AFBBCB class to other classes. Remember that, while terms like 'shuffling' and 'moving to' are common in latent class interpretation, the models are not sequential or related. They are independent solutions that sorted the same students differently. Yet again, it appears that the most important features of the solution are identical across models but there are many different ways to sort students that are equally appropriate, in statistical terms.

Table 5-35

Classes identified by modal responses from LTA and mLTA of the Newton 3 testlet

LTA	% Pre	% Post	mLTA	% Pre	% Post
EEEEAA	6	38	EEEEAA	6	39
AFBBCB	49	17	AFBBCB	41	13
EEEECB	6	25	EEEECB	6	25
AF**CB	10	6	A*BBCB	25	11
AfJjCB	9	2	A*JjCB	10	2
			AfeeCB	8	5
******	4	6	******	5	7
AbbBCB	17	7			

The LTA transition parameters in Table 5-36 indicate some amount of staged learning. Students who began in the correct class were 80% likely to stay in that class and those in the hybrid class were 40% likely to stay in that class or 40% likely to move to the correct class. Students that began the semester with the common misconception were about equally likely to stay, move to the hybrid, or move to the correct class. Very few students moved backwards along that chain. The 'other' classes were most likely to move to the correct or hybrid classes.

Table 5-36

Probabilities of transitioning into posttest classes (column) given class membership at pretest (row) on the Newton 3 testlet.

		1 AfJjCB	2 ******	3 AFBBCB	4 AbbBCB	5 EEEEAA	6 EEEECB	7 AF**CB
1	AfJjCB	0.077	0.077	0.173	0.049	0.358	0.202	0.064
2	******	0.048	0.310	0.092	0.065	0.267	0.212	0.007
3	AFBBCB	0.010	0.036	0.237	0.037	0.386	0.238	0.056
4	AbbBCB	0.006	0.071	0.137	0.229	0.221	0.301	0.037
5	EEEEAA	0.000	0.063	0.000	0.007	0.815	0.092	0.023
6	EEEECB	0.019	0.034	0.018	0.024	0.432	0.452	0.022
7	AF**CB	0.018	0.025	0.098	0.000	0.440	0.225	0.195

The multi-group transition probabilities in Table 5-37 show that the BOTH and TUTORIAL treatment groups had more favorable transitions than the NOT BOTH group. Specifically, those students who started with the common misconception were more likely to transition into the dual conception or to the correct answer class. The students who began the semester with the dual conception were fairly similar across groups. The BOTH and TUTORIAL group students that began the semester in the 'other' classes were more likely to transfer to the dual conception and correct answer classes. The results for the BOTH and TUTORIAL groups appear similar. However, looking specifically at the students in the pretest A*jjCB class, those in the BOTH group were much more likely than those in the TUTORIAL group to move the correct answer class.

Table 5-37

Probabilities of transitioning into posttest classes (column) given class membership at pretest (row) on the Newton 3 testlet, for each treatment group.

NOT BOTH

		1 A*JjCB	2 EEEEAA	3 EEEECB	4 AfeeCB	5 AFBBCB	6 ******	7 A*BBCB
1	A*JjCB	0.229	0.206	0.120	0.078	0.237	0.000	0.131
2	EEEEAA	0.045	0.770	0.105	0.043	0.000	0.000	0.131
3	EEEECB	0.041	0.360	0.433	0.088	0.033	0.045	0.000
4	AfeeCB	0.000	0.244	0.286	0.234	0.236	0.000	0.000
5	AFBBCB	0.025	0.322	0.124	0.070	0.400	0.030	0.029
6	******	0.097	0.115	0.233	0.000	0.054	0.434	0.067
7	A*BBCB	0.002	0.199	0.224	0.052	0.140	0.054	0.328

BOTH

		1 A*JjCB	2 EEEEAA	3 EEEECB	4 AfeeCB	5 AFBBCB	6 ******	7 A*BBCB
1	A*JjCB	0.045	0.462	0.213	0.013	0.125	0.076	0.066
2	EEEEAA	0.000	0.941	0.020	0.000	0.019	0.020	0.000
3	EEEECB	0.017	0.529	0.411	0.027	0.016	0.000	0.000
4	AfeeCB	0.035	0.578	0.180	0.177	0.021	0.000	0.009
5	AFBBCB	0.015	0.466	0.211	0.051	0.192	0.019	0.046
6	******	0.000	0.320	0.244	0.029	0.032	0.252	0.122
7	A*BBCB	0.001	0.334	0.238	0.035	0.075	0.054	0.264

TUTORIAL

		1 A*JjCB	2 EEEEAA	3 EEEECB	4 AfeeCB	5 AFBBCB	6 ******	7 A*BBCB
1	A*JjCB	0.086	0.206	0.265	0.000	0.020	0.244	0.179
2	EEEEAA	0.012	0.733	0.141	0.010	0.000	0.104	0.000
3	EEEECB	0.000	0.407	0.445	0.026	0.000	0.063	0.059
4	AfeeCB	0.000	0.338	0.286	0.096	0.010	0.073	0.108
5	AFBBCB	0.002	0.438	0.247	0.058	0.134	0.046	0.074
6	******	0.053	0.185	0.214	0.000	0.149	0.275	0.124
7	A*BBCB	0.011	0.280	0.394	0.016	0.024	0.082	0.194

Energy

The energy testlet, summarized in Table 5-38, explores student understanding of potential and kinetic energy by asking them about the speed and kinetic energy of a sled at the bottom of a frictionless sledding hill. The correct concept is that, in the absence of friction, the only variable of consequence is the height of the hill. A child adds gravitational potential energy to the sled by carrying it to the top of the hill. During the slide, gravity converts that potential energy into kinetic energy. Many students believe that the steepness of the hill is a factor because in a real, frictional world, steep hills allow for greater acceleration, which feels like speed. In a frictionless world, however, the steepness and acceleration are independent of the final speed of the sled.

Table 5-38

The Energy testlet of the FMCE, in abbreviated form.

Testlet Stem:	An image shows a child pulling a sled up to the top of a hill. The explains that after a frictionless slide down the hill, the sled has a speed v and kinetic energy E.
Item	Asks students to predict…
44	…the **speed** of the sled at the bottom of a **steeper** hill.
45	…the **kinetic energy** of the sled at the bottom of a **steeper** hill.
46	…the **speed** of the sled at the bottom of a **taller, less steep** hill.
47	…the **kinetic energy** of the sled at the bottom of a **taller, less steep** hill.
Response	
A	Greater than the original hill
B	The same as the original hill
C	Lesser than the original hill
D	Not enough information
J	None are correct

The most common responses at pretest and posttest are listed in Table 5-39. Note that for the most response patterns, item pairs 44/45 and 46/47 receive the same response. Kinetic

energy and velocity are positively related by the equation KE = ½(mass)(velocity)^2, so if one quantity is larger the other must also be larger. Those students who answered differently on these pairs may not know the formula or that the concepts are so closely related. The correct answer set is BBAA, indicating that the steeper hill generates the same speed but the taller hill generates a greater sled speed. Students with the response pattern AADD, AACC, AABB, and AAAA, all believe that steepness causes greater speed but have different ideas about the importance of the height of the hill. The AAAA students believe that both height and steepness contribute to a greater velocity while the AACC students believe that steepness is the only factor that matters. AADD and AABB appear to believe that the two features of the hill interact in a compensatory way. The AABB students assume that the height and steepness cancel out while the AADD students believe the two are compensatory but do not assume the effects are equal. Note the BBBB response pattern that accounts for 2-6% of student responses, though it is difficult to interpret.

Table 5-39

Ten most common response patterns on the Energy testlet at pretest and posttest.

PRETEST		POSTTESST	
BBAA	0.13	BBAA	0.41
AADD	0.14	AADD	0.07
AACC	0.11	AACC	0.08
ABCB	0.04	ABCB	0.02
AAAA	0.03	AAAA	0.05
BBBB	0.02	BBBB	0.06
AABB	0.02	AABB	0.04
ABCA	0.02	ABCA	0.01
ABDA	0.02	ABDA	0.01
ABDB	0.02		
		ABAA	0.01
Total	0.55		0.76

Table 5-40 shows the modeling summary of the Energy testlet. Note that, though this testlet had more missing data than the others due to its position at the end of the instrument, the estimation procedure used all available information and did not have any extra difficulty converging on solutions. Each analysis included at least 3000 students. This situation is adequate for the current study, which is a proof of concept of the mLTA method, but would not be adequate for the purposes of making causal inferences. This testlet may have greater bias than other analyses given that the students who did not finish the instrument were more likely to be slower, less motivated, or less careful students.

The latent class analyses at both pretest and posttest suffered from model identification issues, with condition numbers in the 10^{-16} and 10^{-15} range. The combined LCA and LTA models did not have the same problems, possibly because of the greater sample size. The latent transition analysis models had entropy levels below the 0.8 cutoff, though the mLTA model did not have the same problem. The Energy testlet had the smallest contingency table of the FMCE and tended to produce more converged replications than other testlets. The mLTA model with fit statistically significantly better than its constrained counterpart (190, 84 df, $p < .001$), though the BIC of the constrained model was smaller.

Table 5-40

Model estimation summary for the Energy testlet.

Model	Classes	Rep.	Conv.	LL	BIC	Entropy	Error
Pre	6	Yes	20	-15705	32250	.796	
	7	Yes	19	-15628	32237	.811	NPD e-16, IM
	8	Yes	18	-15575	32272	.790	NPD e-12
Post	7	No	20	-10281	21528	.917	NPD e-13, IM
	8	Yes	15	-10203	21512	.920	NPD e-15, IM
	9	No	16	-10155	21555	.917	
Both	7	No	16	-26666	54388	.825	IM
	8	Yes	17	-26493	54194	.856	IM
	9	Yes	11	-26393	54146	.855	IM
LTA	7	Yes	19	-25776	52900	.754	
	8	Yes	19	-25575	52757	.754	
	9	Yes	8	-25446	52778	.750	NPD e-14, IM
mLTA	6	Yes	19	-30685	63079	.812	IM
	7	Yes	17	-30408	62987	.817	IM
	8	No	15	-30195	63067	.821	NPD e-15, IM
Constrained	7	Yes	17	-30503	62471	.807	IM

The solutions to the LCA models are shown in Table 5-41 in the form of modal responses. Each solution includes BBAA, AADD, AACC and AABB classes, which are discussed above. In the pretest data, the AAAA solution appears to be combined with other response patterns to form the Aa** class. In each of the models, the students who answered inconsistently across ‘speed’ and ‘kinetic energy’ were placed in an AB** class or in the **** class. The BBBB class, which accounts for 8% of the posttest group and 6% of students overall, is distinct from the other classes but their response is difficult to interpret. Overall, the combined data solution appears to be an appropriate combination of the pretest and posttest models. The muddled pretest classes split into the AAAA, BBBB, and **** classes in the combined solution.

Table 5-41

Classes identified by modal responses in pretest, posttest, and combined LCAs of the Energy testlet

Pre LCA	%	Post LCA	%	Combined	%
BBAA	17	BBAA	45	BBAA	29
AADD	17	AADD	7	AADD	14
AACC	11	AACC	8	AACC	10
aaBB	9	AABB	6	AABB	6
AB*b	18	aB**	10	AB**	13
*ddD	3				
Aa**	26				
		****	8	****	14
		AAAA	7	AAAA	7
		BBBB	8	BBBB	6

The LTA class solution in 5-42 appears to be identical to the combined LCA solution. The mLTA, which was best represented by seven classes rather than eight, appears to have combined the AABB class with the mysterious BBBB class and formed the abBB class. This may not be an appropriate combination in terms of student thinking. It seems unlikely that the two response sets indicate the same mental model of potential and kinetic energy. Other than that feature, the mLTA solution seems sensible. The most classes are distinct from one another, with an entropy value of 0.8, and they match the most common response patterns described above.

Table 5-42

Classes identified in LTA and mLTA of Energy testlet results.

LTA	% Pre	% Post	mLTA	% Pre	% Post
BBAA	15	46	BBAA	16	48
AADD	20	8	AADD	20	8
AACC	13	7	AACC	15	8
AABB	7	8			
AB**	19	7	AB**	17	6
****	20	11	****	18	10
AAAA	4	5	AAAA	5	6
BBBB	2	7			
			abBB	9	13

The transition probabilities of the LTA model are shown in Table 5-43 and show a pattern very different from the other testlets. In the case of energy concepts, there does not appear to be any halfway or hybrid state. Students appear to have either gone to the correct answer at posttest or stuck with their responses from the pretest. In each row, the greatest probabilities are the correct class (column 4) or the same class (on the diagonal). All other transitions have a probability of 0.10 or less. The students in the AAAA class were more likely than other pretest classes to transfer into the correct answer group, and less likely to stick with the same class, but no other group was particularly likely to transfer into AAAA. The results seem to indicate that learning on this topic does not happen in the same kinds of stages as the other concepts. However, keep in mind that there may be more to student conceptions of kinetic energy than can be captured using this set of items.

Table 5-43

Probabilities of transitioning into posttest classes (column) given class membership at pretest (row) on the Energy testlet.

		1 ****	2 BBBB	3 AB**	4 BBAA	5 AACC	6 AABB	7 AADD	8 AAAA
1	****	0.340	0.079	0.075	0.261	0.073	0.079	0.045	0.048
2	BBBB	0.057	0.180	0.062	0.498	0.020	0.065	0.082	0.037
3	AB**	0.028	0.081	0.137	0.417	0.082	0.103	0.102	0.051
4	BBAA	0.039	0.058	0.004	0.866	0.004	0.005	0.006	0.017
5	AACC	0.085	0.054	0.081	0.354	0.210	0.079	0.058	0.079
6	AABB	0.090	0.062	0.063	0.350	0.051	0.253	0.070	0.060
7	AADD	0.063	0.060	0.071	0.472	0.037	0.053	0.186	0.059
8	AAAA	0.020	0.083	0.027	0.550	0.090	0.109	0.011	0.110

The transition parameters of the mLTA analysis in Table 5-44 show that the BOTH and TUTORIAL treatments were more effective than the NOT BOTH condition. In all rows, students in those groups were more likely to transition to the correct answer and less likely to stay the same. In this solution, the probability of changing from one incorrect class to another was sometimes greater than 0.10. This may be due to the fact that there are fewer classes in the model, so transitions are not spread so thin across potential outcomes. In terms of comparing treatment groups in a large scale, randomized, controlled study the mLTA model shows that the groups are different. Unfortunately, the results do not provide any more detailed information beyond which groups were more successful.

Table 5-44

Probabilities of transitioning into posttest classes (column) given class membership at pretest (row) on the Energy testlet, for each treatment group.

NOT BOTH

		1 ****	2 AACC	3 AB**	4 AADD	5 abBB	6 AAAA	7 BBAA
1	****	0.276	0.180	0.075	0.059	0.205	0.059	0.146
2	AACC	0.037	0.311	0.076	0.092	0.175	0.096	0.212
3	AB**	0.000	0.140	0.154	0.192	0.192	0.072	0.252
4	AADD	0.088	0.028	0.085	0.264	0.088	0.056	0.392
5	abBB	0.000	0.005	0.143	0.116	0.320	0.071	0.345
6	AAAA	0.064	0.116	0.013	0.019	0.171	0.260	0.357
7	BBAA	0.024	0.000	0.006	0.000	0.070	0.008	0.893

BOTH

		1 ****	2 AACC	3 AB**	4 AADD	5 abBB	6 AAAA	7 BBAA
1	****	0.332	0.072	0.059	0.038	0.167	0.046	0.285
2	AACC	0.046	0.200	0.141	0.054	0.127	0.060	0.372
3	AB**	0.000	0.089	0.136	0.083	0.149	0.044	0.499
4	AADD	0.039	0.064	0.062	0.188	0.095	0.075	0.478
5	abBB	0.116	0.066	0.046	0.097	0.263	0.076	0.336
6	AAAA	0.000	0.151	0.083	0.024	0.170	0.000	0.571
7	BBAA	0.028	0.000	0.000	0.004	0.051	0.023	0.894

TUTORIAL

		1 ****	2 AACC	3 AB**	4 AADD	5 abBB	6 AAAA	7 BBAA
1	****	0.324	0.038	0.088	0.016	0.046	0.071	0.417
2	AACC	0.129	0.145	0.031	0.033	0.142	0.074	0.445
3	AB**	0.109	0.080	0.052	0.084	0.133	0.080	0.462
4	AADD	0.069	0.015	0.083	0.132	0.118	0.045	0.539
5	abBB	0.072	0.035	0.011	0.099	0.273	0.038	0.473
6	AAAA	0.083	0.076	0.000	0.000	0.112	0.055	0.674
7	BBAA	0.047	0.015	0.019	0.022	0.041	0.015	0.840

CHAPTER 6

DISCUSSION

This chapter provides a discussion of the results presented in Chapter 5. It begins by addressing each of the research questions individually, describing trends in the LCA and LTA solutions and identifying which testlets defy those trends. The next section describes where the study fits in the greater context of cognitive diagnostic assessment. This section includes recommendations on using latent class modeling for diagnostic assessments and recommendations for further research.

What latent classes are present in the data set? How do the proportions vary across the two time points and the treatment conditions?

The latent classes, presented in Chapter 5 in the form of modal response patterns, seem to represent different versions of the same conceptual knowledge. All of the testlets produced a correct answer class, where the modal responses indicate Newtonian thinking. Each testlet solution also included a common misconception class with students that provided the responses predicted by the cognitive research that informed the assessment design. Two of the six testlets produced classes representing hybrid conceptions, where pieces of correct and incorrect reasoning merge into a distinct way of thinking. Two other testlets had dual conceptions, where students possess both the correct and incorrect versions of the concept and apply both at different times. The testlets that allowed for dual conceptions are those with sub-testlets where students are required to switch contexts. All testlets also generated a mix of 'other' classes with varying degrees of correctness. There are two testlets that stand out as having different class structures: Energy and Force Graphs.

Unlike the other FMCE testlets, the Energy items do not elicit a single most popular misconception. There is some anecdotal evidence that there are many more versions of the potential-kinetic energy misconception than previously thought (Wittmann, 2015). The formal research on student understanding of energy, summarized by Ding (2007), identifies a large number of incorrect ideas that interfere with students answering items correctly. While it seems that the research lacks a systematic taxonomy of ideas, it is clear that student ideas about energy may be too complicated for the FMCE testlet to capture. A complete assessment would require, at the very least, items about a taller hill with the same steepness. The Energy testlet should be flagged for further research and revision.

The Energy testlet has more incorrect classes that show coherent reasoning than the other testlets but lacks an identifiable hybrid or dual conception class. Table 6-1 reproduces the LTA results presented earlier in Table 5-42. The BBAA class knows that height rather than steepness determines the speed and energy of a sled at the bottom of a hill. The AACC students believe the opposite, that height is irrelevant but only steepness matters. These two groups represent the Newtonian and common misconception, respectively. The AABB students believe that steepness is important but assume that steepness and height are compensatory and will cancel out. The AADD students believe that steepness is important, but are not sure how to react when both steepness and height change simultaneously. These two groups are not as easy to label. They both demonstrate the common, erroneous belief that steepness matters but they also believe that height has some impact. It seems that these students have one correct piece of the concept and one incorrect piece, but that they do not form a new, unique idea. The distinction between a hybrid class and a half-correct class is unclear and may not be a distinction that can be made in some cases.

Table 6-1

Modal response patterns for the eight class LTA solution of the Energy testlet (excerpt from Table 5-42)

LTA	% Pre	% Post
BBAA	15	46
AADD	20	8
AACC	13	7
AABB	7	8
AB**	19	7
****	20	11
AAAA	4	5
BBBB	2	7

The energy testlet solution highlights that latent class modeling and testlet structure can vary widely by the specifics of the content area. The classes here represent the interaction of four pieces of information:

- Kinetic energy and velocity are directly related (if not, AB**)
- Steepness is irrelevant to final velocity
- Height is directly related to final velocity
- Steepness and height, if they both affect velocity, directly compensate

It seems that latent class modeling is ideal for this situation, but it may be that the single latent variable is not the best model to use. There are many smaller categorical pieces of knowledge that are activated and assembled to form coherent sets of responses. The potential-kinetic energy concept is likely best modeled using categorical latent variables, though the four-item testlet may not be capable of drawing out the full range of student responses. It lacks a question about a taller hill with the same steepness, at the very least. Regardless of the ideal energy assessment,

this is an excellent example of how assessment structure and measurement model need to be tailored to specific content.

The Force Graphs testlet is another exception to the structure, though in contrast to the Energy testlet, there were fewer identifiable classes than expected. Students overwhelmingly selected the correct EAEBBGE response or the misconception ACBHDF* response, which accounted for 60% of all students at pretest and posttest. The AC***F* and Eae**f* classes appear to be somewhat incorrect and somewhat correct, respectively, but do not have strong item response probabilities. It is impossible to make any confident statement about what students in those classes were thinking when they took the FMCE. This model solution is surprising because the Force Graphs items could be used to express a very similar hybrid conception to that in the Acceleration Graphs testlet. The response pattern ECEHDF* would represent students that have the force-velocity confusion but still know that constant velocity means zero force. That response pattern does not appear in any model solutions nor the list of common response sets. It is unclear from these results why the class appears for the AG testlet but not the FG testlet when the two should be similar. It may be due to features of the items or differences in how students conceptualize force and acceleration.

Table 6-2

Class modal response sets for the six class mLTA solution of the Force Graphs testlet (excerpt from Table 5-21)

mLTA	% Pre	% Post
EAEBBGE	8	36
ACBHDF*	52	29
AC***F*	24	8
Eae**f*	3	17
c****f*	9	8
hA**BgE	3	3

While the testlet solutions match expectations fairly well by including the major classes, but also included collections of ‘other’ classes that capture the variety of student responses. The Force Sled testlet solution in Table 6-3 shows the three types of ‘other’ classes that appear in Chapter 5. The ***** class truly represents the leftovers of the data, those that did not fit in any other class or even fit with each other to form a coherent group. The B**f* students show a

single defining feature, the B response for Item 1, that sets them apart from the other classes. These are students that answered the first item correctly but did not give the correct BDFFB response pattern, else they would have been assigned to the correct class. These types of classes can be labelled 'other correct' or 'other incorrect' because they have a defining correct or incorrect feature but lack homogeneity. The AB*G* class is an 'other incorrect' class, though it has more overall homogeneity. These students are not members of ABCGE or ABFGB but still demonstrate the essential features of the common misconception. The AB*G* class could be labelled as a 'coalition' class because the full set of item response probabilities reveal that it is a union of ABGGA, ABCGA, and ABEGC responses.

Table 6-3

Class modal response sets for the six class multigroup latent transition analysis of the Force Sled testlet

mLTA	% Pre	% Post
BDFFB	12	41
ABCGE	48	23
ABFGB	19	18
B**f*	5	7
*****	2	1
AB*G*	14	9

The solutions to slightly different models (the same testlet at different time points, with one fewer latent class, compared to previous results [Davenport, 2013], etc.) seemed *somewhat* different. Latent class modeling supposes that subgroups exist within the population, and therefore appear in any large sample from that population. However, because of the way the models are structured and estimated, fluctuations across samples will produce slightly different solutions. This idea of 'surface similarity' is a recurring theme in this discussion and it

highlights the notoriously 'fuzzy' nature of latent class modeling. In all the FMCE analyses presented here—and in many models that were tested but not presented—the important classes always appeared in every solution and accounted for 50-80% students in the sample. Meanwhile, the 'other' classes differed across solutions, meaning that 20-50% of students were assigned to groups that may have been poorly defined.

Differences across solutions take two forms. First, the small 'other' classes have different modal response patterns. These classes are necessary for the model to sort every individual in the sample and, as such, are sensitive to small variations from sample to sample. Second, while each class is best described by its modal parameters, its full definition includes dozens of parameters. These non-modal parameters can vary from solution to solution, allowing a class to absorb—or discriminate against—one of the less popular response patterns. The 'surface similarity' phenomenon may be a weakness in the LCA approach to FMCE scoring, but may also be an advantage. It is the flexibility of LCA that allows it to capture both well-defined mental states and the hybrid, dual, or 'other' states.

What does LTA reveal about conceptual change in physics over a semester of instruction?

The latent transition analyses show that students do change conceptual knowledge over a semester of instruction. The majority of students begins each semester with a dominant misconception or as members of 'other incorrect' classes. At the end of the semester, the correct or hybrid classes represent a narrow majority of students. While this is a good sign in terms of learning, physics instructors hope that *all* students leave introductory courses with an understanding of these fundamental physics concepts. The information that LTAs provide about *how* students learn may be useful for improving physics instruction and reaching the goal of universal Newtonian thinking.

The dominant pattern of transitions, referred to here as staged learning, is displayed in Figure 6-1 using example values from the seven-class LTA solution of the Acceleration Graphs testlet. Each numerical value is a transition probability, conditional on membership in the pretest class (where each arrow begins), or the probability of staying in the same class. The small sixth and seventh classes were omitted, as were any transition probabilities smaller than 0.1. The resulting diagram is a description of how conceptual knowledge of acceleration graphs changes for the majority of students. Notice that the paths all tend to move from incorrect states to correct states. Approximately 57% of students beginning with the common misconception move to the correct answer class, though 17% stay in the misconception class, and 13% move to the hybrid class. Those beginning in the hybrid or 'other correct' class are overwhelmingly likely to move to the correct classes. The overall flow of learning is from incorrect to correct where some students stop along the way in a hybrid state.

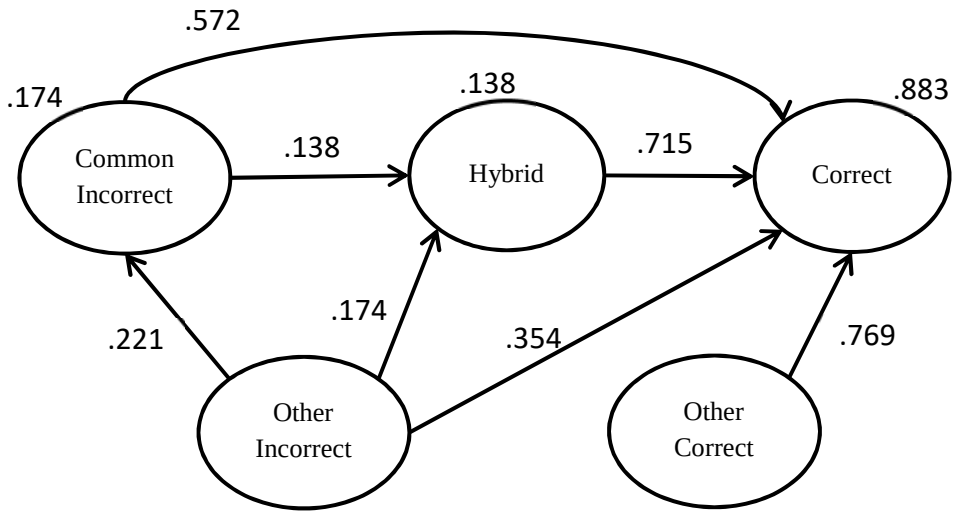


Figure 6-1. LTA Transition probabilities for the largest classes of the Acceleration Graphs testlet.

Four of the six testlets (FS, RD, AG, and N3) have transition structures similar to Figure 6-1. The Force Graphs and Energy testlets do not have the same structure as they do not seem to have hybrid or dual conceptions. The Energy transitions appear sensible for a class structure with no intermediary states, where the only probabilities greater than 0.1 are moving to the correct response class or staying in the same class. The LTA of the Force Graphs testlet shows a much more complicated structure due to the 'other incorrect' and 'other correct' classes. In this case, the flow of student knowledge from incorrect to correct has many intermediate states.

One question about conceptual change that remains is whether students with common misconceptions are better or worse off than students in the 'other' classes. Specifically, which initial state has higher probabilities of forward motion through the typical transition structure. The question was raised by more than one of the physics teacher interviewed for a preliminary study described in Chapter 2. One line of reasoning is that students in 'other' classes know so little about physics that they do not answer coherently, and so are 'further behind' than their peers. The other thought is that students in 'other' classes are not burdened by the naïve conceptions that are resistant to change and make learning physics concepts difficult.

The selection of transition parameters presented in Table 6-4 shows that the answer to this question depends on the testlet and the concept. In the Force Sled testlet, students that start the semester in 'other' classes seem to have a distinct advantage over common misconception students: They are more likely to transition to the correct class by the end of the semester. In the Force Graphs testlet, it is the common misconception students have the clear advantage. In the Newton's Third Law testlet, the 'other' classes are much more likely to move to the dual conception class, likely a stepping stone towards the correct answer class. Meanwhile, in the Acceleration Graphs testlet, the two 'other' groups presented have very different probabilities of

moving to the correct class, one greater and one less than that of the misconception class. The Energy testlet and the Newton 3 testlet both have skeptical classes (AADD and AfJjCB, respectively) where students use the ‘not enough information’ response. These classes appear to have a higher probability of moving to the correct response class.

Table 6-4

A selection of transition parameters, conditional on pretest membership in selected incorrect classes

Testlet	Class	Incorrect	Hybrid	Correct
FS				
	ABCGE	0.397	0.145	0.250
	AB*Gc	0.147	0.255	0.354
	ABgGA	0.086	0.193	0.394
FG				
	ACBHDF*	0.333		0.351
	AC***F*	0.358		0.195
	C****F*	0.234		0.132
AG				
	EGBFA	0.174	0.138	0.572
	E***a	0.218	0.147	0.715
	Eg**A	0.221	0.174	0.417
N3				
	AFBBCB	0.237	0.238	0.386
	AfJjCB	0.173	0.202	0.358
	AbbBCB	0.137	0.301	0.221
	AF**CB	0.098	0.440	0.225
E				
	AACC	.210		0.354
	AADD	.186		0.472
	AB**	.102		0.417

Note: RD testlet excluded due to the large number of hybrid/incorrect classes.

Of all the transitions under discussion, the most important to content area experts may be the transition that represents wholesale change. This is the change that happens when students move from the dominant misconception directly to the correct response. Table 6-5 shows the probability of wholesale change, conditional on pretest misconception class membership, for each testlet. The probabilities range from 25% to 57%, values lower than instructors would

hope. The values in Table 6-5 are intended as descriptive statistics only, as they are single parameters in a much larger model solution. The values should not be compared against one another in an inferential way, as they come from testlets covering different topics, with different structures, and varying models.

Table 6-5

Transition probabilities to the correct class, conditional on membership in the dominant misconception class

Testlet	Transition Probability
Force Sled	0.25
Reverse Direction	0.26
Force Graphs	0.35
Acceleration Graphs	0.57
Newton 3	0.39
Energy	0.35

Can multi-group LTA detect significant differences between treatment groups?

The most important result of the current study is the successful estimation of multi-group latent transition analyses with item responses as categorical indicators. Separate transition parameters were calculated for each of the treatment groups, describing student transitions across a semester of instruction. Unfortunately, it is unclear whether the transitions were statistically significantly different. The likelihood ratio test showed that the free models were statistically significantly better fitting than the constrained models. However the BICs of the constrained models were smaller than the multigroup models, indicating better fit. A search of the literature has not revealed any clear way to reconcile these differences, nor an explanation of why the inconsistency occurred in each of the study's model comparisons. The discussion that follows

assumes some difference across groups, given the observations from the transition tables, a precedent set by Collins and Lanza (2010).

The transitions of the NOT BOTH group, an approximation of a control condition, were substantially different from the transitions of the BOTH and TUTORIAL groups. The intervention groups had transition parameters that showed greater 'motion' from incorrect to correct mental models, as described in the previous section. Currently, there is no established method for testing contrasts of specific parameters across groups, because each parameter uses a different reference value. That method may be found or developed in future research, but this study is limited to wholesale tests of group differences. The multiple regressions presented in Chapter 5 show that posttest scores across the three groups are significantly different after controlling for pretest scores. This study shows that mLTA models probably provide at least as much information as conventional scoring and inferential tests.

Does mLTA provide more information than a raw score comparison?

The results presented in Chapter 5 show that categorical scoring provides as much information as a raw score regression, in terms of comparing group performance. However, the goal of diagnostic testing is to provide information beyond which students performed better than others. Unfortunately, the mLTA transitions do not appear to provide any insight above what was already provided by the conventional analyses. In all cases, the BOTH group and the TUTORIAL group had more favorable transitions than the NOT BOTH group. More students in those groups transitioned to the correct class, more transferred to and from the hybrid and dual classes, and fewer students stayed as members of incorrect classes. But the mLTA transitions provided little insight into how students learned differently over a semester of each treatment

condition. So a combination of regular LTA models and conventional inferential tests would have provided the same amount of information as the mLTA models.

Despite this result, the mLTA model does have the capability to provide detailed information about student learning across curricula. The method itself, which provides specific transition probabilities for different treatment groups, appears to work well in most cases. The lack of useful, comparative inferences is not for a lack of fine-grained diagnostic information. It may be that a comprehensive method for interpreting transition parameters is required. It may also be that there are no particularly interesting comparisons in this particular scenario. The two intervention curricula may simply be more effective than the NOT BOTH condition. In either case, further research is required. Analytical methods may find hidden gems of diagnostic information in the transition tables of Chapter 5. Studies that investigate other curricula, or use other conceptual assessments, may find a number of diagnostically relevant differences. The method itself appears to be solid, but its usefulness to diagnostic assessment is unproven until further research can be conducted.

Do the answers to these questions vary across the testlets? Can differing results be attributed to learning differences, testlet structure, or modeling issues?

The structure of classes varied across testlets. Force Sled and Acceleration Graphs had hybrid classes, Reverse Direction and Newton 3 had dual classes, and the Force Graphs and Energy testlets did not appear to have either. This difference may be due to the structure of the testlet with respect to the concept being targeted. The dual classes appeared in testlets that were broken into sub-testlets with slightly different item stems. The context switching gave some students the opportunity to demonstrate both the correct and incorrect mental. The two testlets that generated hybrid classes do not appear to have any special characteristics, simply asking

about each possible permutation of motion. It seems likely that the Energy testlet does not have a hybrid class because, as described above, the content has a specific structure where conceptual elements can be combined in different ways without producing hybrid mental models, which are characterized by distinctness and 'oddness.'

It is possible that the distinction I make between hybrid and half correct conceptions is not meaningful. Consider two examples from two FMCE testlets. The ABFGB students in the Force Sled testlet know that forces need to oppose motion to slow an object, but they confuse velocity and force when in constant motion or speeding up. The AACC and AADD students in the Energy testlet believe incorrectly that steepness matters to velocity at the bottom of the hill but also believe correctly that the height matters to the velocity. The former was deemed a hybrid conception because the response is inconsistent to the outside observer but maintains conceptual consistency within itself. The latter example was termed a 'half-correct' response because it demonstrates that students believe both attributes of the hill are important when, in reality, only one is important. It could just as easily be argued that AACC and AADD are like the ABFGB students who have half the concept correct.

The differences across testlets described thus far appear to be differences in the relationship between the cognitive model and the test structure (see *Figure 6-2* below, reprinted from Chapter 1). The appearance of hybrid or dual conceptions depends on their presence in the minds of the population *and how those ideas map to the items and responses.* On the other hand, at least one difference across testlets appears to depend on the relationship between test structure and measurement model. The failure of the Reverse Direction testlet to converge is not a problem of mapping concepts to item responses. In fact, the response patterns of the RD testlet are the most readily interpretable, where most students gave some version of GDB or AAA to

each of the three sets of items. It is likely that the convergence issues stem from the number of combinations that can be generated: GDB AAA AAA, GDB GDB AAA, GDB AAA GDB, and so on. Berzofsky, Beimer, and Kalsbeek (2014) showed that identification issues can arise from violations of local independence. The RD testlet has three sub-testlets where item responses are more related to the item subgroup than to the other items in the testlet. Whether it was intradependence issues or simply that the model required far too many classes and parameters to converge, the Reverse Direction testlet had issues because of the link between test structure and measurement modeling.

A scheme for developing effective and valid assessments

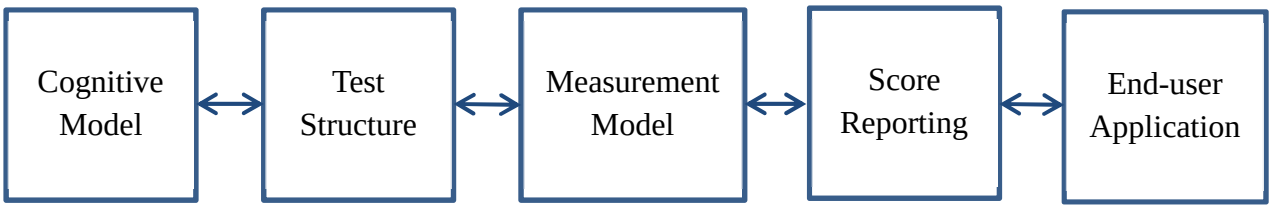


Figure 6-2. A scheme for developing effective and valid assessments.

The model identification issues also varied across testlets. The Reverse Direction testlet had consistent problems across models, while the Newton 3 testlet problems were specific to the mLTA model and the Force Sled testlet had problems with all solutions except those with the best number of classes. Collins and Lanza (2010) stated that models with more classes and more parameters tend to have more identification issues. Muthen (2007b) explained that identification is slightly more complicated. If a solution has a class with very few students, then there is not enough information to determine the parameters for that particular class. So while classes generally get smaller as the number of classes increases, identifiability issues depend on *how*

those classes splinter off. For example, in the Reverse Direction testlet, very few students gave the GDB AAA GDB response pattern but those students were very distinct from the others. Assuming a sufficient number of classes, the GDB AAA GDB students will split off from the rest and form their own class, too small to estimate properly but too distinct to combine with. Here it appears that the mental model, the test structure, and the measurement model all interact.

In terms of the main research questions, whether mLTA can be used to find differences between class transitions using a concept inventory, the testlets seemed similar. With the exception of the RD testlet, the mLTA and constrained models showed that transitions varied across the treatment groups. The transition probabilities themselves showed that the BOTH and TUTORIAL groups were more successful at changing student knowledge than the NOT BOTH group. All testlets were also similar in that the transitions provided little diagnostic information.

Greater Context

Exploring categorical measurement

One of the secondary goals of this study is to examine the properties of categorical scoring models used to measure categorical constructs. The results shed some light on typical FMCE results, which are driven by a mismatch of categorical constructs and continuous modeling. As an example, consider Cronbach's α, calculated to be .939 for pretest FMCE data and .956 for posttest data. Values close to 1 indicate that scale items correlate highly with one another. In other words, if a student answers one item correctly, then they are likely to answer more items correctly. This derived value is often used by content area researchers as the sole criterion for determining test reliability when it is actually a measure of internal consistency, just one aspect of reliability. If a test had poor internal consistency, perhaps because items are driven

by two continuous, unrelated constructs, different students might get the same overall score even though they had different values of the underlying traits. Similarly, two students with the same values of the target construct might receive different scores because of different levels on the second, interfering construct. This scenario would likely be represented by a low value of Cronbach's α.

When the underlying trait is categorical, rather than continuous, conventional measures of internal consistency become inflated. In the example of the Reverse Direction testlet, students with the correct concept answered all nine items correctly while students with the common misconception got all nine items incorrect. In fact, only about 40% of students scored anything between zero or nine points on the RD testlet. As a result, the very strong α value of 0.931 appears to indicate a very consistent continuous scale when, in reality, the mathematics are driven to an inflated value by the categorical underpinnings of the testlet. Other surface features found by conventional analyses, such as strong item response theory discrimination parameters (Davenport, 2013), are likely influenced by the categorical-continuous mismatch.

To get a sense of internal consistency within a categorical framework, it is more appropriate to consider class separation and homogeneity. The fundamental question of reliability is whether an instrument produces the same values for two individuals with the same level of the underlying construct. In the continuous case, this means that students with the same value of the target construct should produce similar scores. In the categorical case, reliability means that students with the same mental state should be sorted into the same classes. Lack of internal consistency threatens reliability in the same way as in continuous scoring, but it takes on different forms (e.g., items that provide no information about classification).

As an example, consider the Reverse Direction combined pretest/posttest data classes presented in Table 6-6. The two largest classes are distinguished by different responses on all nine items. Since the item response probabilities are very strong for these two classes, we can infer that students in those classes have categorically different beliefs about the forces during a change in direction. The situation is not as clear for the 'other' classes. There is uncertainty what the 'other' classes say about student thinking and uncertainty about whether the students in each 'other' class really share the same ideas.

Table 6-6

The combined LCA results for the Reverse Direction testlet

Class Modal Response	% of Students
AAA AAA AAA	20
GDB GDB GDB	28
GDB GDB AaA	8
FDb fDB *DB	7
GDB GDB cDf	7
gDa gDA gDa	7
gDb AAA AAA	6
d *d* *d*	4
G*b gaB *ab	4
*Da *DA AdA	6
D AAA *Db	4

The last two classes inTable 6-6, *Da *DA AdA and *D* AAA *Db, share many features. It is mainly Item 12 (the fifth of nine items) that differentiates the two classes. This item asks students about the force on a coin at the peak of its motion. The former class has a high probability of responding with D to this item and the latter class has a high probability of responding with A. If the distinction between the two classes relies on Item 12, students with the same construct might end up in either class due to fluctuations in their responses to that one item.

This kind of narrow division between classes might causea low entropy value. At the same time if the Item 12 parameter values remain highly discriminating, the entropy could remain high while students might still switch classes on the back of a single item. This highlights how the strength of latent class modeling, that it can classify on specific features and does not require all features to differ, can also be a weakness. The idea of internal consistency needs to be re-conceptualized for an entirely categorical model.

Researchers rely on entropy to provide information about cross-classification and class separation because it is often taken to represent a distinctness of classes.The Reverse Direction example has an entropy value of .909, which is considered to be strong. The example from the table above shows that entropy is a valuable tool but does not tell the whole story. If the parameters regarding item 12 are great enough, students will never be cross-classified between the two classes. At the same time, perhaps they should be cross-classified because their responses share so many features.

In both the classical and LCA measurement frameworks, an errant response can change a students' score away from a true score. In the continuous case, the errant response gives the student extra points or fewer points. In the categorical case, the errant response may or may not cause the student to be classified incorrectly. This kind of measurement error is unavoidable but should be minimized whenever possible.

Multidimensionality, where ideas unrelated to the target construct can interfere with inference, manifests differently in latent class models. An example would be the Energy testlet where height and/or steepness might contribute to the velocity of the sled. The two distinct ideas are evaluated together, which might be a threat to internal consistency in a conventional model, but in LCA simply results in a larger number of classes. If the model suffered from the large

number of classes, one option would be to model the multidimensionality by creating two latent variables for each of height and steepness. Threats to reliability, such as extra dimensions, might result in cross-classification, might be dismissed by low item response probabilities, or absorbed into the class structure itself. Either way, the conventional measure of internal consistency will appear strong because it is driven by the all correct and all incorrect students.

The mismatch between continuous scores and categorical constructs can also explain the features of score distributions (reproduced as Figure 6-3 below). In the pre-post raw score scatterplots presented in Chapter 4, two groups of students were visible: those that began with low scores and ended with a wide range of scores and those that began above the 'floor' and almost unanimously hit the 'ceiling.' The plots show strong positive skew and a ceiling effect. This result is problematic because the items are too difficult for most pretest students and too easy for approximately one third of the posttest students. In the language of item response theory, the test does not provide enough information about low end pretest students or high end posttest students. A typical recommendation would be to add more difficult items and more easy items to the instrument.

Adding difficulty is problematic for concept inventories that target simple, core ideas. When the questions are as simple as 'what force causes the object to move to the right with a steadily increasing velocity' there are only a few options for increasing difficulty. Items could ask about changing forces and non-steady increases in velocity, but that content is out of the domain of introductory physics (usually taught as a part of a second year mechanics course using differential equations). More difficult items could use specific contexts that are tricky for students, as in the 'pushing' items of the Newton 3 testlet, but that strategy carries the danger of adding construct-irrelevant variance. The other option is to change the wording of the item to

use more difficult language which, again, is likely to add construct-irrelevant variance. Similar problems occur with adding easier items to measure pretest students more effectively. When the items are already as simple as ‘what force would keep an object moving at the same velocity,’ there is no way to make the item easier.

Skew, floor, and ceiling effects are not necessarily problematic for criterion-referenced instruments. For example, when designing a criterion-referenced assessment that targets a single cut score, designers select discriminating items that are near the cut score in terms of difficulty. The items provide a large amount of information at the specific level to ensure that students are identified as being above or below the cut score. Such instruments are poorly equipped to describe how far above or below the cut score students are but that is not particularly problematic for placement tests. In the case of the FMCE, the designers aimed a large number of items at the proficiency level where students start to ‘get’ Newton’s laws of motion. It seems that the FMCE provides information at a very specific level, where students ‘either get it or they don’t.’ Students bunch at the low end and high end and are spread widely across the middle range of scores, because the FMCE provides considerable information about students who are right in the midst of learning the concepts.

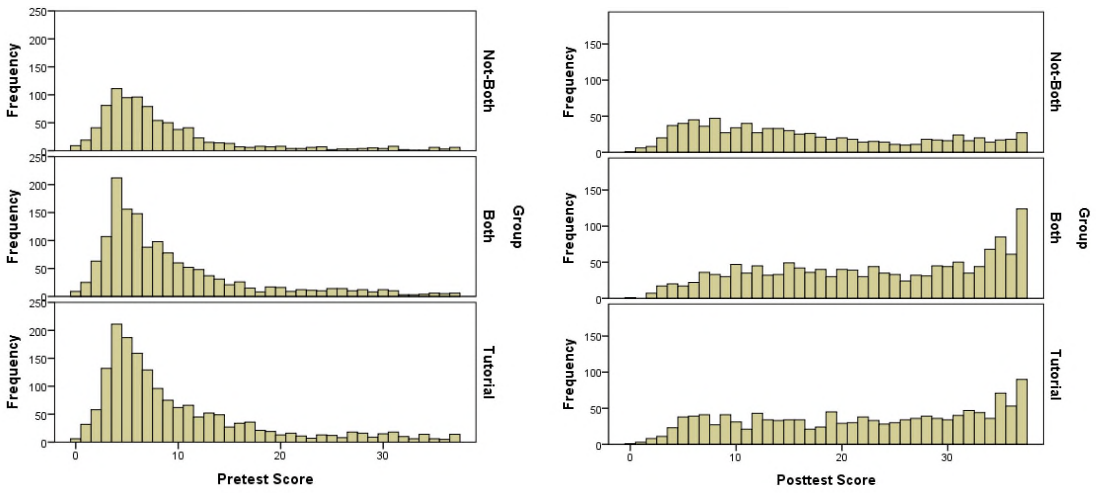


Figure 6-3. Conventional score distributions of pretest (left) and posttest (right) students from three treatment groups.

While the distributions appear consistent with the FMCE as a criterion-referenced instrument, these results are problematic for the way the FMCE is used in practice. Proficiency tests with specific cut-scores are used to separate individuals for the purposes of placement or licensure. The FMCE is used to compare the impact of teaching methods on samples of students, which is most effective with information about students across the proficiency range. Floor and ceiling effects restrict the variability of scores, creating bias in correlations and regression coefficients. In this sense, the FMCE fails to meet the criteria set by the design scheme in Figure 6-2 when conventional scoring is used. The link between measurement model and end-user application is faulty.

The chain of quality instrument design can be repaired by modeling the FMCE in a categorical framework. Unlike a conventional model, the LCA model allows for classes that give all correct and all incorrect responses. Additionally, it can classify the students that answer incorrectly into meaningfully different classes. The continuous scores show that the FMCE is

targeted at posttest students who are right at the point of understanding difficult concepts, showing how scores are spread widely across a narrow conceptual range. However, it is the LCA model that explains the intermediary states that drive the spread of scores in that range. In some cases, students have intermediate knowledge states. In other cases, students have correct concepts for some testlets but not others. Either way, the categorical model explains the features seen in the continuous model while providing more detailed information.

The appropriateness and usefulness of latent class modeling

A major premise of this dissertation is that the FMCE is best described by categorical measurement modeling and that, while conventional methods work well enough for some purposes, the continuous models obscure the diagnostic information available in the data. The study raised many questions, such as why the Force Graphs testlet did not produce any hybrid conceptual states, which should be addressed by further research. The question that remains to be discussed here is whether latent class modeling is appropriate and worthwhile in the context of large-scale intervention research.

First, it is important to consider the expertise and resources required to perform the mLTA analyses. Latent class analysis is widely known among psychometricians and methodologists but is still esoteric among content area researchers. Latent transition analysis is less common, but is still within the skill set of most methodologists. The mLTA models were more sophisticated than either LCA or LTA models. Any research team would need an experienced methodologist to use mLTA models. The study would also need to use an assessment similar to the FMCE, with a testlet structure targeting specific concepts that take different categorical states within the minds of students. The researchers would also need access to a computer with many processing cores, given the huge amount of computation time.

A second issue to consider is the difficulty of interpreting latent class results when using multiple choice responses as indicators. Latent class modeling is notoriously 'fuzzy.' Selecting the best number of classes is a process that lacks rigorous criteria, particularly in situations where the Vuong-Lo-Mendell-Rubin and the Bootstrap Likelihood Ratio Test provide no usable information. The classes are defined by a large number of parameters which can be summarized using the modal responses, though that obscures the full profile of each class. The measurement invariance across group and across time remains an unanswered question, casting doubts on inferences. In the analyses reported here, there were numerous instances of solutions that were statistically different but substantively similar. Perhaps most concerning are the conflicting results regarding the free and constrained models. Until the conflicts are resolved and a better way of testing varying parameters is found, results of actual studies would remain inconclusive. The overall fuzziness of latent class modeling is a factor that needs to be considered when designing a study.

The final consideration for recommending this method is that the mLTA models provided little diagnostic information beyond the LTA and regression models. The comparisons of mLTA models against constrained models were successful and readily interpretable. These tests, statistically significant for all testlets, indicated that transition parameters were different across treatment groups. The transitions themselves show that the BOTH group and the TUTORIAL group had more favorable transitions than the NOT BOTH group. Students in those groups were more likely to transition to the correct class and less likely to stay with the common misconception. On the other hand, without a test to compare specific transitions, it is difficult to make specific claims. Most LTA results are analyzed by interpreting transition parameters descriptively (e.g. Collins & Lanza, 2010). In that spirit, I can say that mLTA models find

differences across groups and provide descriptions of how those groups learn. While future applications of the mLTA method may be more diagnostically fruitful, it is difficult to recommend the method after a single application that provided few diagnostic conclusions.

Overall, it is difficult to recommend mLTA, with multiple choice responses as indicators, for analyzing randomized controlled trial results. It requires many resources, has 'fuzzy' results, and did not provide specific diagnostic results in this context. However, the mLTA models *did* work in the sense that they converged and provided interpretable solutions. The mLTA models might be more useful in other contexts. This study was successful as a proof of concept. While it may not be feasible for content area researchers to take this method and immediately apply it, with further studies and applications, the mLTA models could be a useful addition to the cognitive diagnostic toolbox.

Future Research

This study showed that mLTA models can be used with conceptual surveys and controlled trials. However, it is just the beginning of a long line of research to (1) answer questions raised during this study, (2) to evaluate categorical measurement modeling as a practice, and (3) to eventually use categorical scoring to answer some fundamental questions about conceptual knowledge.

One problem that needs to be addressed is whether the latent classes match up with student thinking and whether interpretations of the classes are accurate. The correct classes and the common misconceptions classes were easy to identify, as those conceptual states were identified by Thornton and Sokoloff (1998) and written into the design of the FMCE. The hybrid and dual classes have not been explored qualitatively. Interview studies with students

who provided hybrid and dual response patterns could confirm the existence of these states and describe them more accurately. The 'other' classes appear to be mixes of students with different ideas, but interviews might provide some insight into what students in these classes believe. A qualitative study of student thinking is a necessary step before using LCA to perform content area research using the FMCE.

Another problem that needs attention is the question of measurement invariance. One of the results of this study was that across time points, the latent class models did not have the same number of classes, though the most important and populated classes were the same. Across groups, the latent class structure was similar but not exactly the same. Models that freely estimated the item response parameters across groups had identification issues and gave conflicting results. Clearly, this is not ideal and needs to be addressed. The essential question of measurement invariance, in this context, is whether students with the same conceptual knowledge will be assigned to the same class regardless of their group membership. For the main classes (correct, common misconception, and hybrid) the answer to the invariance question appears to be affirmative. For 'other' classes, the question is remains unresolved. Qualitative research with students responding with 'other' response patterns may shed some light on the issue. Another possible exploration is picking a specific set of response patterns and calculating how each is classified by different model solutions. Simulation studies could answer questions about measurement invariance in the context of conceptual assessments. They can, with incrementally increasing variance, find the point at which students tend to be classified differently. Simulations could help tease apart the differences in solutions that are due to real differences in class structure and those due to the fluctuations across samples.

The next problem that needs to be explored is that of class structure and model identification. Collins and Lanza (2010) warn that models with more classes and more parameters tend to have identification problems, while Muthen (2007b) explains that identification problems can occur within a single class if that class has fewer members than relevant parameters. On the other hand, Berzofsky and Beimer (2014) found that violations of conditional independence cause identification problems. For the Reverse Direction testlet, model identification was a major problem though it is not clear which explanation for non-invariance is most relevant—or if the three explanations are separable. The RD testlet may violate conditional independence with its structure of three sub-testlets. The latent class models require a large number of classes to account for performance on each sub-testlet (GDB GDB GDB, GDB GDB AAA, GDB AAA AAA, et cetera). With so many classes, the models have an enormous number of parameters and some classes have very few members. The violations of conditional dependence may require more parameters and classes to describe than can be identified, relating the three explanations. Simulation studies would shed some light on these model identification issues and might be able to identify the threshold of complexity where identification issues appear.

To continue the evaluation of categorical scoring and the mLTA model, the methods presented here need to be applied to other instruments and to other treatments. The current study was intended as a proof-of-concept, and produced mixed results. The models do converge and do provide information about student learning across groups, though it seems that they provide little diagnostic data beyond that provided by LTA in combination with raw score analyses. Another trial with a different data set, preferably with different treatments, may provide more diagnostic data. A trial using a different instrument would help evaluate whether categorical

scoring has potential for more general use. Tests with an additional instrument might also shed some insight into the invariance and identification issues described above.

If those studies prove fruitful, the next step would be to design an instrument with latent class modeling in mind. The challenge is to write items and responses that access conceptual knowledge and are also structured in a way that the latent class models will function properly. The simulation studies and trials would provide information about how to design testlets. Thus far, it appears that the nine-item testlet with three sub-testlets will not function properly. Another challenge to constructing such an instrument would be finding topics where the target conceptual knowledge does appear to be categorical with hybrid or dual states and can be explored with a set of 5-8 multiple choice items.

Finally, if future evaluations show that latent class modeling allows for valid and reliable inferences, it could be used to explore the stability of conceptual knowledge. A test-retest study using the FMCE (Davenport, 2008) showed that students tend to get the same continuous score after four weeks of non-instruction—but that students change many of their answers from one testing occasion to the next. The hybrid and dual conceptions illustrate the flexible and unpredictable nature of knowledge after some amount of instruction. It is likely that students in the test-retest study accessed different versions of their conceptual knowledge and provided substantively different response patterns. Latent transition analyses could provide information about how the prevalence of class shifts with no instruction. This study would provide a description of the stability of knowledge itself. The analysis could be extended to a mover-stayer model that may be able to predict which students have more stable knowledge in the absence of instruction.

As noted earlier, internal consistency is just one aspect of reliability. Another aspect is temporal consistency, whether the same individual would get the same score at two different time points with no instruction in between. The idea of temporal consistency is interesting because the test-retest studies that are used to evaluate it are affected both by the random error of measurement *and* by the flexibility of knowledge itself. Students with misconceptions that have recently learned the correct conception are often left in an ambiguous dual state. Ask a student with one semester of experience with physics about the forces on a coin during a coin flip and they may access the correct or the incorrect version of the knowledge. Ask that same student the same question a week later, they may access the other version of the concept and provide different answers. As of right now, the magnitude of that effect is unclear. Latent transition analyses with conceptual test may, by exploring knowledge stability, provide a more complete picture of reliability itself.

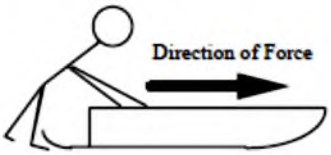
Appendix A: The Force and Motion Conceptual Evaluation

FORCE AND MOTION CONCEPTUAL EVALUATION

Directions: Answer questions 1-47 in spaces on the answer sheet. Be sure your name is on the answer sheet. Answer question 46a also on the answer sheet. Hand in the questions and the answer sheet.

A sled on ice moves in the ways described in questions 1-7 below. *Friction is so small that it can be ignored.* A person wearing spiked shoes standing on the ice can apply a force to the sled and push it along the ice. Choose the one force (A through G) which would **keep the sled moving** as described in each statement below.

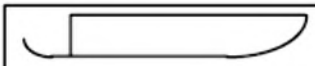
You may use a choice more than once or not at all but choose only one answer for each blank. If you think that none is correct, answer choice **J**.



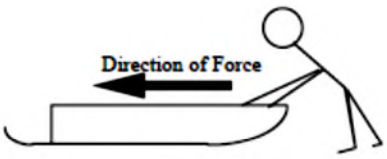
A. The force is toward the **right** and is **increasing** in strength (magnitude).

B. The force is toward the **right** and is of **constant** strength (magnitude).

C. The force is toward the **right** and is **decreasing** in strength (magnitude).



D. No applied force is needed



E. The force is toward the **left** and is **decreasing** in strength (magnitude).

F. The force is toward the **left** and is of **constant** strength (magnitude).

G. The force is toward the **left** and is **increasing** in strength (magnitude).

___1. Which force would keep the sled moving toward the right and speeding up at a steady rate (constant acceleration)?

___2. Which force would keep the sled moving toward the right at a steady (constant) velocity?

___3. The sled is moving toward the right. Which force would slow it down at a steady rate (constant acceleration)?

___4. Which force would keep the sled moving toward the left and speeding up at a steady rate (constant acceleration)?

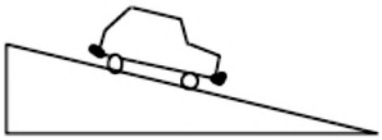
___5. The sled was started from rest and pushed until it reached a steady (constant) velocity toward the right. Which force would keep the sled moving at this velocity?

___6. The sled is slowing down at a steady rate and has an acceleration to the right. Which force would account for this motion?

___7. The sled is moving toward the left. Which force would slow it down at a steady rate (constant acceleration)?

Figure A- 1. The first page of the FMCE, the Force Sled testlet.

Questions 8-10 refer to a toy car which is given a quick push so that it rolls up an inclined ramp. After it is released, it rolls up, reaches its highest point and rolls back down again. *Friction is so small it can be ignored.*



Use one of the following choices (**A** through **G**) to indicate the **net force** acting on the car for each of the cases described below. Answer choice **J** if you think that none is correct.

(A) Net constant force down ramp
(B) Net increasing force down ramp
(C) Net decreasing force down ramp
(D) Net force zero
(E) Net constant force up ramp
(F) Net increasing force up ramp
(G) Net decreasing force up ramp

_____8. The car is moving up the ramp after it is released.
_____9. The car is at its highest point.
_____10. The car is moving down the ramp.

Questions 11-13 refer to a coin which is tossed straight up into the air. After it is released it moves upward, reaches its highest point and falls back down again. Use one of the following choices (**A** through **G**) to indicate the force acting on the coin for each of the cases described below. Answer choice **J** if you think that none is correct. **Ignore any effects of air resistance.**

A. The force is **down** and constant.
B. The force is **down** and increasing
C. The force is **down** and decreasing
D. The force is zero.
E. The force is **up** and constant.
F. The force is **up** and increasing
G. The force is **up** and decreasing

_____11. The coin is moving upward after it is released.
_____12. The coin is at its highest point.
_____13. The coin is moving downward.

Figure A- 2. The second page of the FMCE, two parts of the Reverse Direction testlet.

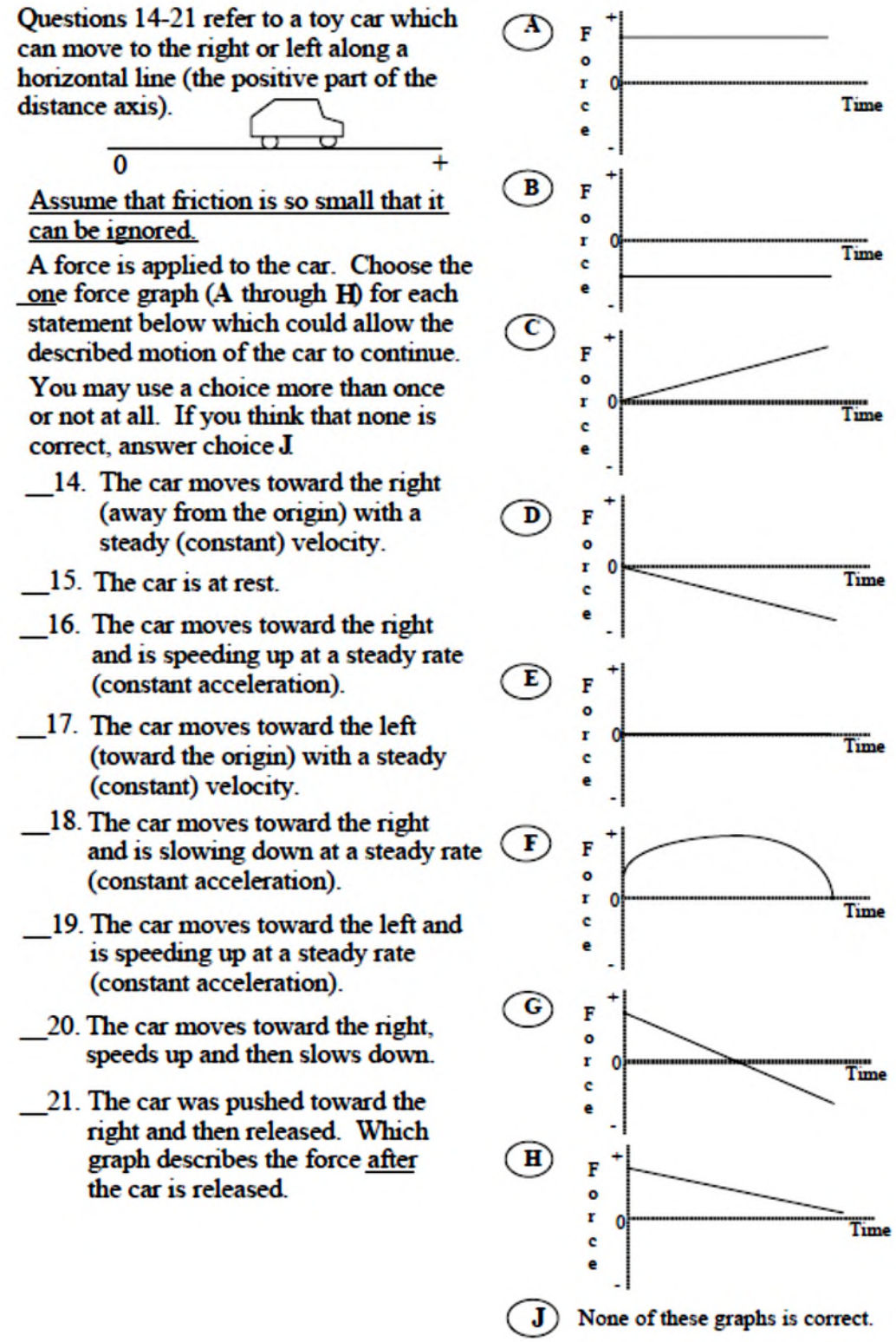
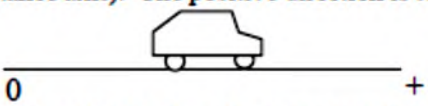


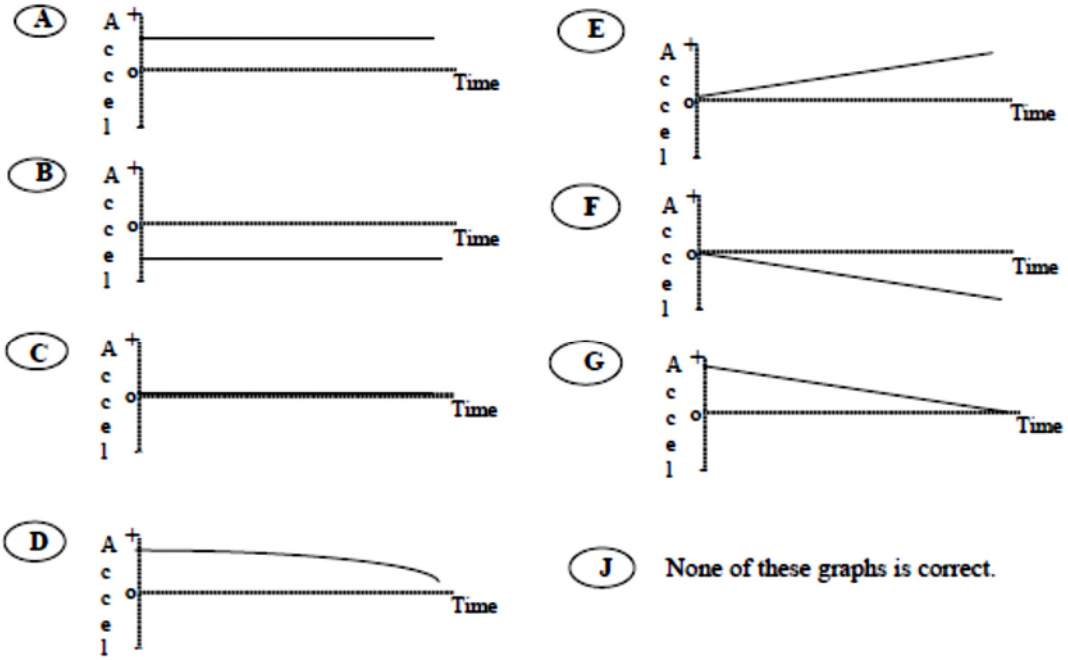
Figure A- 3. The third page of the FMCE, Force Graphs testlet.

Questions 22-26 refer to a toy car which can move to the right or left on a horizontal surface along a straight line (the + distance axis). The positive direction is to the right.



Different motions of the car are described below. Choose the letter (**A** to **G**) of the **acceleration-time** graph which corresponds to the motion of the car described in each statement.

You may use a choice more than once or not at all. If you think that none is correct, answer choice **J**.



_____22. The car moves toward the right (away from the origin), speeding up at a steady rate.

_____23. The car moves toward the right, slowing down at a steady rate.

_____24. The car moves toward the left (toward the origin) at a constant velocity.

_____25. The car moves toward the left, speeding up at a steady rate.

_____26. The car moves toward the right at a constant velocity.

Figure A- 4. The fourth page of the FMCE, the Acceleration Graphs testlet.

Questions 27-29 refer to a coin that is tossed straight up into the air. After it is released it moves upward, reaches its highest point and falls back down again. Use one of the following choices (A through G) to indicate the acceleration of the coin during each of the stages of the coin's motion described below. Take **up** to be the **positive** direction. Answer choice **J** if you think that none is correct.

A. The acceleration is in the negative direction and constant.
B. The acceleration is in the negative direction and increasing
C. The acceleration is in the negative direction and decreasing
D. The acceleration is zero.
E. The acceleration is in the positive direction and constant.
F. The acceleration is in the positive direction and increasing
G. The acceleration is in the positive direction and decreasing

___27. The coin is moving upward after it is released.
___28. The coin is at its highest point.
___29. The coin is moving downward.

Questions 30-34 refer to collisions between a car and trucks. For each description of a collision (30-34) below, choose the one answer from the possibilities **A** though **J** that best describes the forces between the car and the truck.

A. The truck exerts a greater amount of force on the car than the car exerts on the truck.
B. The car exerts a greater amount of force on the truck than the truck exerts on the car.
C. Neither exerts a force on the other; the car gets smashed simply because it is in the way of the truck.
D. The truck exerts a force on the car but the car doesn't exert a force on the truck.
E. The truck exerts the same amount of force on the car as the car exerts on the truck.
F. Not enough information is given to pick one of the answers above.
J. None of the answers above describes the situation correctly.

*In questions 30 through 32 the truck is **much heavier** than the car .*

_____30. They are both moving at the same speed when they collide. Which choice describes the forces?
_____31. The car is moving much faster than the heavier truck when they collide. Which choice describes the forces?
_____32. The heavier truck is standing still when the car hits it. Which choice describes the forces?

Figure A- 5. The fifth page of the FMCE, the first half of the Newton Three testlet.

In questions 33 and 34 the truck is a small pickup and is the same weight as the car.

_____33. Both the truck and the car are moving at the same speed when they collide. Which choice describes the forces?

_____34. The truck is standing still when the car hits it. Which choice describes the forces?

Questions 35-38 refer to a large truck which breaks down out on the road and receives a push back to town by a small compact car.

Pick one of the choices **A** through **J** below which correctly describes the forces between the car and the truck for each of the descriptions (35-38).

A. The force of the car pushing against the truck is equal to that of the truck pushing back against the car.

B. The force of the car pushing against the truck is less than that of the truck pushing back against the car.

C. The force of the car pushing against the truck is greater than that of the truck pushing back against the car.

D. The car's engine is running so it applies a force as it pushes against the truck, but the truck's engine isn't running so it can't push back with a force against the car.

E. Neither the car nor the truck exert any force on each other. The truck is pushed forward simply because it is in the way of the car.

J. None of these descriptions is correct.

_____35. The car is pushing on the truck, but not hard enough to make the truck move.

_____36. The car, still pushing the truck, is **speeding up** to get to cruising speed.

_____37. The car, still pushing the truck, is at cruising speed and continues to travel at the **same speed**.

_____38. The car, still pushing the truck, is at cruising speed when the truck puts on its brakes and causes the car to **slow down**.

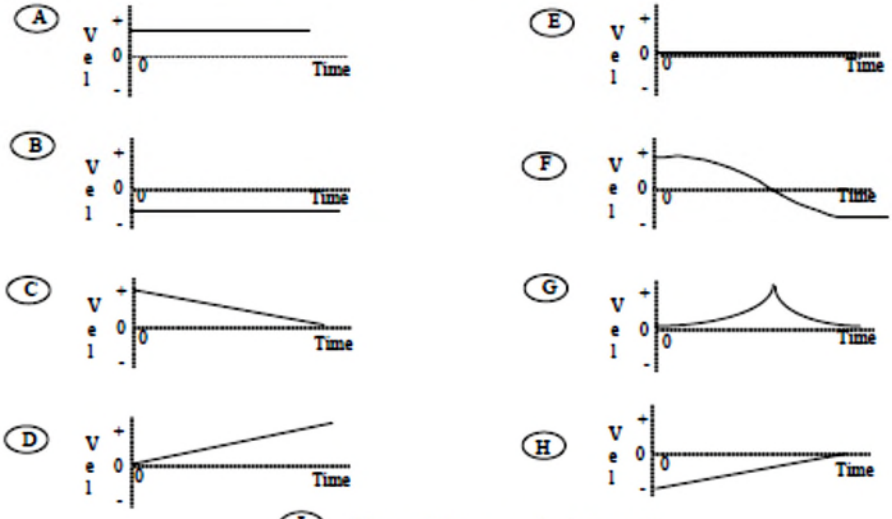
Figure A- 6. The seventh page of the FMCE, the second half of the Newton Three testlet.

___39. Two students sit in identical office chairs facing each other. Bob has a mass of 95 kg, while Jim has a mass of 77 kg. Bob places his bare feet on Jim's knees, as shown to the right. Bob then suddenly pushes outward with his feet, causing both chairs to move. In this situation, while Bob's feet are in contact with Jim's knees,

A. Neither student exerts a force on the other.
B. Bob exerts a force on Jim, but Jim doesn't exert any force on Bob.
C. Each student exerts a force on the other, but Jim exerts the larger force.
D. Each student exerts a force on the other, but Bob exerts the larger force.
E. Each student exerts the same amount of force on the other.
J. None of these answers is correct.

Questions 40-43 refer to a toy car which can move to the right or left along a horizontal line (the positive portion of the distance axis). The positive direction is to the right.

Choose the correct velocity-time graph (**A** - **G**) for each of the following questions. You may use a graph more than once or not at all. If you think that none is correct, answer choice **J**.



J None of these graphs is correct.

___40. Which velocity graph shows the car moving toward the right (away from the origin) at a steady (constant) velocity?
___41. Which velocity graph shows the car reversing direction?
___42. Which velocity graph shows the car moving toward the left (toward the origin) at a steady (constant) velocity?
___43. Which velocity graph shows the car increasing its *speed* at a steady (constant) rate?

Figure A- 7. The seventh page of the FMCE, the Velocity Graphs testlet.

A sled is pulled up to the top of a hill. The sketch above indicates the shape of the hill. At the top of the hill the sled is released from rest and allowed to coast down the hill. At the bottom of the hill the sled has a speed v and a kinetic energy E (the energy due to the sled's motion). Answer the following questions. ***In every case friction and air resistance are so small they can be ignored.***

__44. The sled is pulled up a **steeper** hill of the **same** height as the hill described above. How will the velocity of the sled at the bottom of the hill (after it has slid down) compare to that of the sled at the bottom of the original hill? Choose the best answer below.

A. The speed at the bottom is greater for the steeper hill.
B. The speed at the bottom is the same for both hills.
C. The speed at the bottom is greater for the original hill because the sled travels further.
D. There is not enough information given to say which speed at the bottom is faster.
J. None of these descriptions is correct.

__45. Compare the kinetic energy (energy of motion) of the sled at the bottom for the original hill and the steeper hill in the previous problem. Choose the best answer below.

A. The kinetic energy of the sled at the bottom is greater for the steeper hill.
B. The kinetic energy of the sled at the bottom is the same for both hills.
C. The kinetic energy at the bottom is greater for the original hill.
D. There is not enough information given to say which kinetic energy is greater.
J. None of these descriptions is correct.

__46. The sled is pulled up a **higher** hill that is **less** steep than the original hill described before question 44. How does the speed of the sled at the bottom of the hill (after it has slid down) compare to that of the sled at the bottom of the original hill?

A. The speed at the bottom is greater for the higher but less steep hill than for the original.
B. The speed at the bottom is the same for both hills.
C. The speed at the bottom is greater for the original hill.
D. There is not enough information given to say which speed at the bottom is faster.
J. None of these descriptions is correct.

46a. Describe in words your reasoning in reaching your answer to question 46. (**Answer on the answer sheet** and use as much space as you need)

_47. For the higher hill that is less steep, how does the kinetic energy of the sled at the bottom of the hill after it has slid down compare to that of the original hill?

A. The kinetic energy of the sled at the bottom is greater for the higher but less steep hill.
B. The kinetic energy of the sled at the bottom is the same for both hills.
C. The kinetic energy at the bottom is greater for the original hill.
D. There is not enough information given to say which kinetic energy is greater.
J. None of these descriptions is correct.

Figure A- 8. The eighth page of the FMCE, the Energy testlet.

Appendix B: Continuous Latent Factor Measurement Model Results

One of the challenges for designing conceptual assessments with testlets is that of dimensionality. Conventional scores are often given as summary scores and presented alongside subscale scores for each testlet. Some indicators, such as inter-item correlations and IRT results, suggest that a uni-dimensional model is adequate. The structure of the test, however, suggests that the instrument must be multi-dimensional. In this appendix, I present results from a series of models that use conventional, continuous scoring. The models include exploratory and confirmatory factor analyses, also a bifactor model that explores the space between uni- and multi-dimensionality. To mimic common methodologies, I selected a random half of all cases to use in the EFA procedures and used the other half in the CFA and bifactor procedure.

Exploratory Factor Analysis

Figure B- 1 shows the scree plot of the combined pretest-posttest FMCE data, which is a plot of the eigenvalues of each sequential component of a factor analysis. The data shows two 'elbows' at the fifth and eighth components, indicating that the best solution may be a five or eight factor solution. The result of a parallel analysis is shown in Figure B- 2, in the form of a scree plot where the FMCE eigenvalues are plotted with the eigenvalues of entirely random set of data (O'Connor, 2000). The two lines cross at approximately the 19th component, implicating a substantially larger number of factors. A minimum average partial test (O'Connor, 2000) using tetrachoric correlations indicated that the best solution has five factors. All calculations for these preliminary tests were performed in SPSS Version 21.

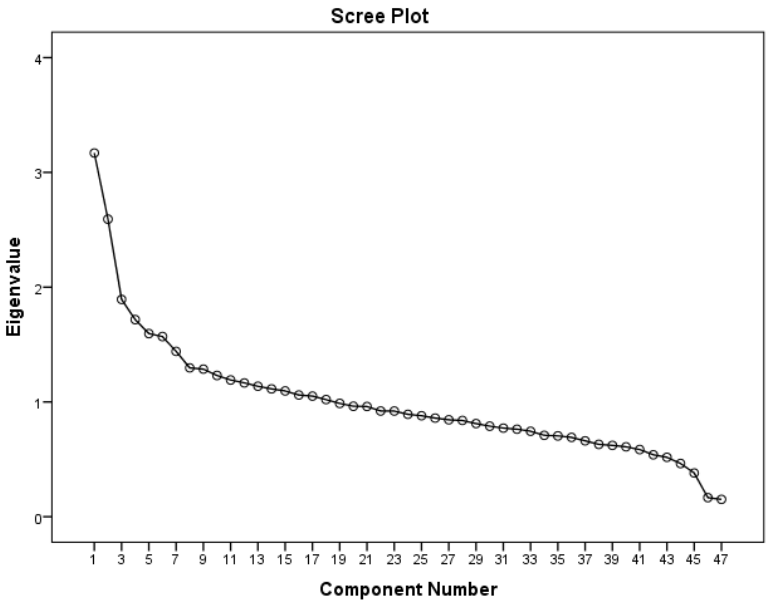


Figure B- 1. Scree plot of the combined pretest and posttest FMCE results. The nineteenth component is the first with an eigenvalue smaller than 1.

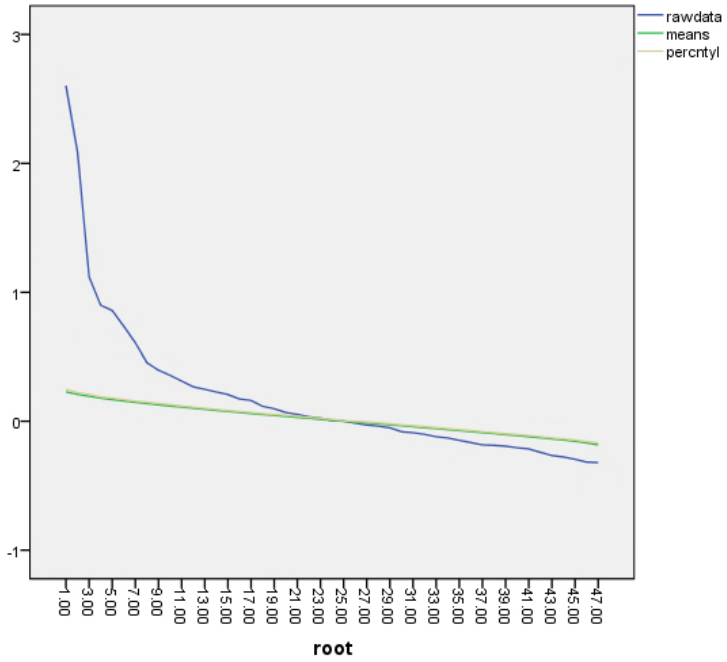


Figure B- 2. The results of a parallel analysis, plotting the observed eigenvalues against those of random data. The two lines cross at approximately the 25th component.

Each test indicated a different number of factors, so I estimated EFA models with increasing numbers of factors to see which had the best model fit. Three separate sets of solutions used pretest data, posttest data, and pretest-posttest combined. Mplus Version 7.11 was used with weighted least mean square estimation, which uses all available data rather than performing deletion or imputation, and the oblique geomin rotation method. Unfortunately, the chi-square difference tests showed that each sequential model fit better than the last, up through 10 factors. I chose to stop estimating models with more than 10 factors as they would be uninterpretable.

With each of the criteria for pointing to a different number of factors, the factor loadings themselves were used to make a final decision. The six factor solution appeared to be the most appropriate for pretest, posttest, and combined data. In each case, moving to the seven factor solution produced extraneous factors that targeted individual items while still leaving two of the testlets loaded on a single factor. The factor loadings from the pretest data are presented in Table B-1. The FMCE appears to have simple structure when six factors are assumed. The third factor combines the Force Graphs and Acceleration Graphs testlets, implying that student performance on those items is strongly related. I will refer to factor three as a Graphs factor. Overall, the solution shows simple structure. The Force Graph items appear to also have some association with the first, Force Sled, factor. The factor correlations are presented in Table B- 2 and are moderately large. The Reverse Direction factor is strongly related to the Force Sled and Graphs factors, as well as the Energy testlet.

Table B- 1

Pattern matrix of factor loadings for a six factor solution using both pretest and posttest data

Item	1	2	3	4	5	6
FMCE1	**0.867***	0.054*	0.141*	-0.014	0.058*	-0.029
FMCE2	**0.663***	0.054*	0.242*	0.053*	0.125*	-0.006
FMCE3	**0.762***	0.001	-0.013	0.005	0.355*	0.038
FMCE4	**0.853***	0.028	0.157*	-0.02	0.060*	-0.027
FMCE7	**0.719***	0.011	-0.032	-0.023	0.368*	0.062*
FMCE8	0.087*	**0.893***	0.006	0.018	-0.011	0.03
FMCE9	0.069*	**0.858***	0.003	0.016	-0.026	0.059*
FMCE10	0.116*	**0.880***	-0.013	0.010	-0.051*	-0.050*
FMCE11	0.054*	**0.882***	0.007	0.085*	-0.009	-0.01
FMCE12	0.026	**0.854***	-0.006	0.099*	0.017	0.027
FMCE13	0.070*	**0.977***	-0.047*	0.023	-0.049*	-0.102*
FMCE14	0.323*	0.026	**0.758***	0.057*	-0.223*	0.022
FMCE16	0.472*	-0.025	**0.608***	0.096*	-0.123*	0.011
FMCE17	0.306*	-0.011	**0.820***	0.017	-0.234*	0.026
FMCE18	0.393*	0.080*	**0.572***	0.026	-0.032	0.034
FMCE19	0.460*	0.025	**0.560***	0.066*	-0.110*	0.013
FMCE20	0.359*	0.196*	0.358*	-0.052*	-0.013	0.064*
FMCE21	0.254*	0.287*	0.311*	-0.027	0.132*	0.103*
FMCE22	0.037	-0.098*	**0.898***	0.034	0.252*	-0.035*
FMCE23	0.033	-0.024	**0.846***	0.029	0.223*	-0.051*
FMCE24	-0.160*	0.067*	**1.005***	-0.040*	0.057*	0.006
FMCE25	0.113*	0.038	**0.698***	-0.004	0.172*	0.032
FMCE26	-0.193*	0.056	**1.010***	-0.027	0.062*	0.008
FMCE27	-0.061*	**0.724***	0.198*	-0.019	0.188*	0.024
FMCE28	-0.081*	**0.735***	0.155*	0.029	0.169*	0.064*
FMCE29	-0.052*	**0.762***	0.164*	-0.028	0.144*	0.002
FMCE30	0.032	0.015	-0.028	**0.953***	-0.048*	0.003
FMCE31	-0.019	0.021	-0.017	**0.898***	0.012	0.001
FMCE32	0.024	-0.033	0.033	**0.960***	0.048*	-0.001
FMCE34	0.045*	0.000	0.035*	**0.932***	0.031*	0.023
FMCE36	-0.044	0.083*	0.024	**0.829***	-0.007	0.024
FMCE38	-0.026	0.053	0.056*	**0.820***	0.007	0.015
FMCE40	-0.005	0.041	0.008	0.027	**0.923***	-0.027
FMCE41	0.043	0.093	0.092*	0.034	**0.466***	0.102*
FMCE42	0.055	0.026	0.085*	0.032	**0.746***	0.015
FMCE43	0.017	-0.021	-0.007	0.063	**0.683***	0.042
FMCE44	0.061*	0.028	0.000	0.125*	0.053*	**0.789***
FMCE45	0.054*	0.008	-0.080*	0.101*	0.033	**0.778***

FMCE46	-0.024	0.014	0.042*	-0.008	-0.036	**0.924***
FMCE47	-0.018	-0.008	0.060*	-0.042*	-0.025	**0.900***

*Statistically significant, p<.05

Table B- 2

Correlations between factors identified by an EFA of combined pretest and posttest FMCE data

	FS	RD	Graphs	N3	VG	E
FS	1					
RD	0.685*	1				
Graphs	0.567*	0.723*	1			
N3	0.461*	0.629*	0.568*	1		
VG	0.164*	0.334*	0.392*	0.211*	1	
E	0.471*	0.600*	0.515*	0.475*	0.278*	1

*Statistically significant, p<.05

The six factor solution loadings for the posttest data are presented in Table B-3. What is most interesting about the posttest factor results is that the Force Graphs items load onto one factor with the Force Sled items instead of with the Acceleration Graphs items. It is possible that, after instruction, performance on those items has more to do with an understanding of forces and less to do with the graphical feature of the items. I will refer to factor 1 as a Forces factor. Overall, the posttest results appear to show simple structure.

Table B- 3

Pattern matrix of factor loadings for a six factor solution using pretest FMCE data

Item	1	2	3	4	5	6
FMCE1	**0.960***	0.005	0.054*	-0.041*	-0.005	-0.033
FMCE2	**0.688***	0.04	0.209*	0.036	0.082*	0.035
FMCE3	**0.827***	-0.023	-0.063*	0.031	0.346*	0.001
FMCE4	**0.969***	-0.027	0.038	-0.051*	0.002	-0.03
FMCE7	**0.793***	-0.002	-0.070*	-0.009	0.349*	0.003
FMCE8	0.280*	**0.718***	0.013	0.071*	0.054*	0.021
FMCE9	0.238*	**0.696***	0.015	0.067*	-0.008	0.012
FMCE10	0.217*	**0.756***	-0.006	0.005	-0.033	-0.115*
FMCE11	0.152*	**0.837***	-0.009	0.072*	0.032	-0.019
FMCE12	0.162*	**0.778***	0.029	0.072*	0.049*	-0.003
FMCE13	0.135	**0.907***	-0.035	-0.023	-0.047*	-0.143*
FMCE14	**0.499***	0.068*	**0.502***	0.082*	-0.316*	0.049*
FMCE16	**0.696***	0.057*	0.344*	-0.007	-0.104*	0.029
FMCE17	**0.499***	0.019	**0.546***	0.080*	-0.307*	0.043
FMCE18	**0.612***	0.102*	0.347*	-0.001	-0.046	0.060*
FMCE19	**0.702***	0.033	0.313*	-0.012	-0.080*	0.039
FMCE20	**0.494***	0.135*	0.256*	0.005	-0.006	0.074*
FMCE21	0.358*	0.230*	0.290*	0.011	0.107*	0.096*
FMCE22	0.145*	-0.073*	**0.885***	-0.006	0.225*	-0.052*
FMCE23	0.152*	-0.041*	**0.825***	0.038*	0.193*	-0.053*
FMCE24	-0.04	0.009	**0.979***	0.006	-0.01	-0.014
FMCE25	0.249*	0.016	**0.681***	-0.006	0.141*	-0.016
FMCE26	-0.056	0.005	**1.014***	-0.011	-0.021	-0.025
FMCE27	-0.011	**0.564***	0.411*	-0.036	0.074*	0.088*
FMCE28	-0.019	**0.629***	0.386*	-0.017	0.039	0.101*
FMCE29	-0.052	**0.671***	0.351*	-0.056*	-0.016	0.013
FMCE30	0.060*	0.018	-0.085*	**0.840***	-0.02	-0.007
FMCE31	0.006	-0.018	-0.012	**0.709***	0.094*	0.011
FMCE32	0.058*	-0.046*	0.034	**0.950***	0.190*	-0.058*
FMCE34	0.118*	-0.013	0.012	**0.929***	0.147*	-0.029
FMCE36	-0.160*	0.263*	-0.009	**0.860***	-0.065*	0.057*
FMCE38	-0.156*	0.236*	0.038	**0.862***	-0.084*	0.042
FMCE40	-0.080*	0.175*	0.018	0.004	**0.889***	0.038
FMCE41	0.055	0.117*	0.160*	-0.004	0.444*	0.077*
FMCE42	0.014	0.135*	0.083*	0.01	**0.724***	0.065*
FMCE43	-0.068	0.116	0.043	0.02	**0.626***	0.021
FMCE44	0.073*	0.000	0.016	0.061*	0.046*	**0.806***
FMCE45	0.058	0.000	-0.027	0.025	0.017	**0.687***

FMCE46	0.009	0.010	0.003	-0.009	-0.035	**0.857***
FMCE47	0.019	0.000	0.024	-0.02	-0.012	**0.781***

*Statistically significant, p<.05

The correlations in Table B- 4show the relationships among posttest factors, which are not as strong as with the pretest data but still moderately strong. Note the very low correlation between the Velocity Graphs factor and the other factors. This is likely due to the VG items being too easy and having little variation. The Reverse Direction factor again has the highest correlations with other factors.

Table B- 4

Correlations between factors identified by an EFA of pretest FMCE data

	Forces	RD	AG	N3	VG	E
Forces	1					
RD	0.618*	1				
AG	0.511*	0.523*	1			
N3	0.394*	0.420*	0.345*	1		
VG	0.032	0.166*	0.253*	-0.043	1	
E	0.464*	0.491*	0.404*	0.332*	0.097*	1

*Statistically significant, p<.05

The next analysis I performed was an EFA using both the pretest and posttest data combined. The factor loadings in Table B- 5 again show simple structure. The fairly clean structure of these factor loadings indicates that the FMCE is somewhat multi-dimensional with dimensions closely associated with the testlets. With this full data set, the Force Graph items were more strongly associated with the Acceleration Graph items than the Force Sled items. I named factor 3 a Graphs factor. The correlations in Table B- 6 illustrate moderate to strong relationships among the first three factors. The strength of these correlations supports the idea that the test is somewhat uni-dimensional.

Table B- 5

Pattern matrix of factor loadings for a six factor solution using postest FMCE data

Item	1	2	3	4	5	6
FMCE1	**0.871***	0.022	0.210*	-0.004	-0.041*	0.007
FMCE2	**0.713***	0.016	0.260*	0.031	0.027	0.048*
FMCE3	**0.764***	0.102*	-0.016	-0.021	0.203*	-0.021
FMCE4	**0.823***	0.017	0.243*	0.004	-0.008	-0.009
FMCE7	**0.701***	0.112*	-0.016	-0.021	0.235*	0.000
FMCE8	0.009	**0.924***	0.061*	0.004	-0.046*	0.058*
FMCE9	-0.008	**0.891***	0.034	0.004	-0.007	0.086*
FMCE10	0.049*	**0.874***	0.059*	-0.007	-0.059*	-0.009
FMCE11	0.025	**0.912***	0.015	0.038*	-0.009	0.033*
FMCE12	-0.045*	**0.913***	-0.039*	0.055*	0.033*	0.071*
FMCE13	0.037	**0.921***	0.013	0.017	-0.016	-0.024
FMCE14	0.289*	-0.012	**0.816***	0.013	-0.039*	0.048*
FMCE16	0.377*	-0.042*	**0.741***	0.073*	-0.058*	-0.009
FMCE17	0.245*	-0.047*	**0.851***	-0.01	-0.02	0.053*
FMCE18	0.311*	0.091*	**0.626***	0.046*	0.073*	0.019
FMCE19	0.332*	0.021	**0.672***	0.050*	-0.006	0.003
FMCE20	0.304*	0.193*	0.349*	-0.023	0.094*	0.057*
FMCE21	0.248*	0.333*	0.237*	-0.002	0.180*	0.097*
FMCE22	0.007	-0.042	**0.689***	0.041*	0.543*	-0.048*
FMCE23	0.016	0.050*	**0.634***	0.006	0.511*	-0.054*
FMCE24	-0.125*	0.041	**0.764***	-0.034*	0.481*	0.002
FMCE25	0.094*	0.060*	**0.573***	0.01	0.381*	0.005
FMCE26	-0.150*	0.030	**0.737***	-0.021	0.492*	0.001
FMCE27	0.033	**0.763***	-0.017	0.01	0.321*	-0.048*
FMCE28	-0.013	**0.750***	-0.032	0.028	0.306*	0.012
FMCE29	0.066*	**0.730***	0.011	-0.009	0.286*	-0.055*
FMCE30	0.031	-0.004	-0.043*	**0.965***	-0.024	0.000
FMCE31	0.034	0.004	-0.045*	**0.907***	0.033	-0.016
FMCE32	0.024	-0.037*	0.022	**0.969***	0.028	0.021
FMCE34	0.032*	0.01	0.012	**0.948***	0.006	0.028
FMCE36	-0.085*	0.136*	0.074*	**0.766***	-0.022	0.01
FMCE38	-0.072*	0.123*	0.093*	**0.755***	0.007	-0.01
FMCE40	-0.011	-0.034	-0.001	-0.001	**0.913***	0.092*
FMCE41	0.043	0.081	0.037	0.015	**0.451***	0.161*
FMCE42	0.073*	-0.012	0.087*	0.009	**0.736***	0.075*
FMCE43	0.021	-0.089	-0.04	0.087*	**0.701***	0.166*
FMCE44	0.201*	0.021	-0.028*	0.078*	0.017	**0.799***
FMCE45	0.198*	-0.043*	-0.064*	0.086*	0.035	**0.820***

FMCE46	-0.031*	0.058*	0.109*	-0.037*	-0.021	**0.922***
FMCE47	-0.029*	0.038	0.113*	-0.056*	-0.003	**0.909***

Table B- 6

Correlations between factors identified by an EFA of posttest FMCE data

	FS	RD	Graphs	N3	VG	E
FS	1					
RD	0.627*	1				
Graphs	0.453*	0.613*	1			
N3	0.345*	0.464*	0.358*	1		
VG	0.323*	0.363*	0.255*	0.253*	1	
E	0.385*	0.484*	0.337*	0.360*	0.327*	1

When a test is split into subscales corresponding to specific factors, the reliability of each factor must be considered. Table B-7 shows the α values for each factor identified in the pretest, posttest, and combined EFAs. It also includes the mean and standard deviations of the subscale scores, calculated with one point per item rather than the scoring template described in Chapter 5. The α values are very strong with the exception of the Velocity Graphs scale which suffers from a lack of variability.

Table B- 7

Means, standard deviations, and Cronbach's alpha for factors identified by pretest, posttest, and combined exploratory factor analyses

Pretest	Items	α	Mean (SD)
Force Sled	1-4, 7	.882	1.76 (1.96)
Reverse Direction	8-13, 27-29	.935	3.22 (3.46)
Graphs	14-19, 22-26	.927	5.27 (4.10)
Newton 3	30-32, 34, 36, 38	.917	2.23 (2.40)
Velocity Graphs	40-43	.694	3.34 (1.05)
Energy	43-47	.835	1.88 (1.62)

Posttest	Items	α	Mean (SD)
Forces	1-4, 7, 14-19	.897	2.82 (3.00)
Reverse Direction	8-13, 27-29	.892	1.65 (2.51)
Accel. Graphs	22-26	.899	1.70 (1.98)
Newton Three	30-32, 34, 36, 38	.839	0.98 (2.66)
Velocity Graphs	40-43	.667	3.20 (1.23)
Energy	43-47	.757	1.32 (2.02)

Combined	Items	α	Mean (SD)
Force Sled	1-4, 7	.885	2.49 (2.06)
Reverse Direction	8-13, 27-29	.931	4.90 (3.55)
Graphs	14-19, 22-26	.914	6.89 (3.71)
Newton Three	30-32, 34, 36, 38	.901	3.57 (2.33)
Velocity Graphs	40-43	.710	3.53 (.923)
Energy	43-47	.861	2.42 (1.64)

Confirmatory Factor Analysis

I used the structure identified in the combined data EFA described above to create confirmatory factor analysis models, removing Items 20 and 21 which were not related strongly to any single factor. The path diagram for the model is shown in Table B-7 and abbreviates the correlation paths among the factors using a circular joint at the center of the diagram. Each latent factor is correlated with each other factor. Note that the model also includes an error

variance parameter for each observed indicator variable, though the error variances were not included in the path diagram to conserve space. The model was run using a weighted least mean square estimation in Mplus that uses as much data as available in each case. The model had a goodness-of-fit chi-square of 7644 with 650 degrees of freedom, which indicates that the model-implied variance-covariance matrix is significantly different from the observed data matrix. The root mean squared error of the estimated model is .052, statistically significantly greater than the preferred .05 value, and had a CFI of .982 and a TLI of .981. These values indicate that the model did not fit the data particularly well, but not particularly poorly.

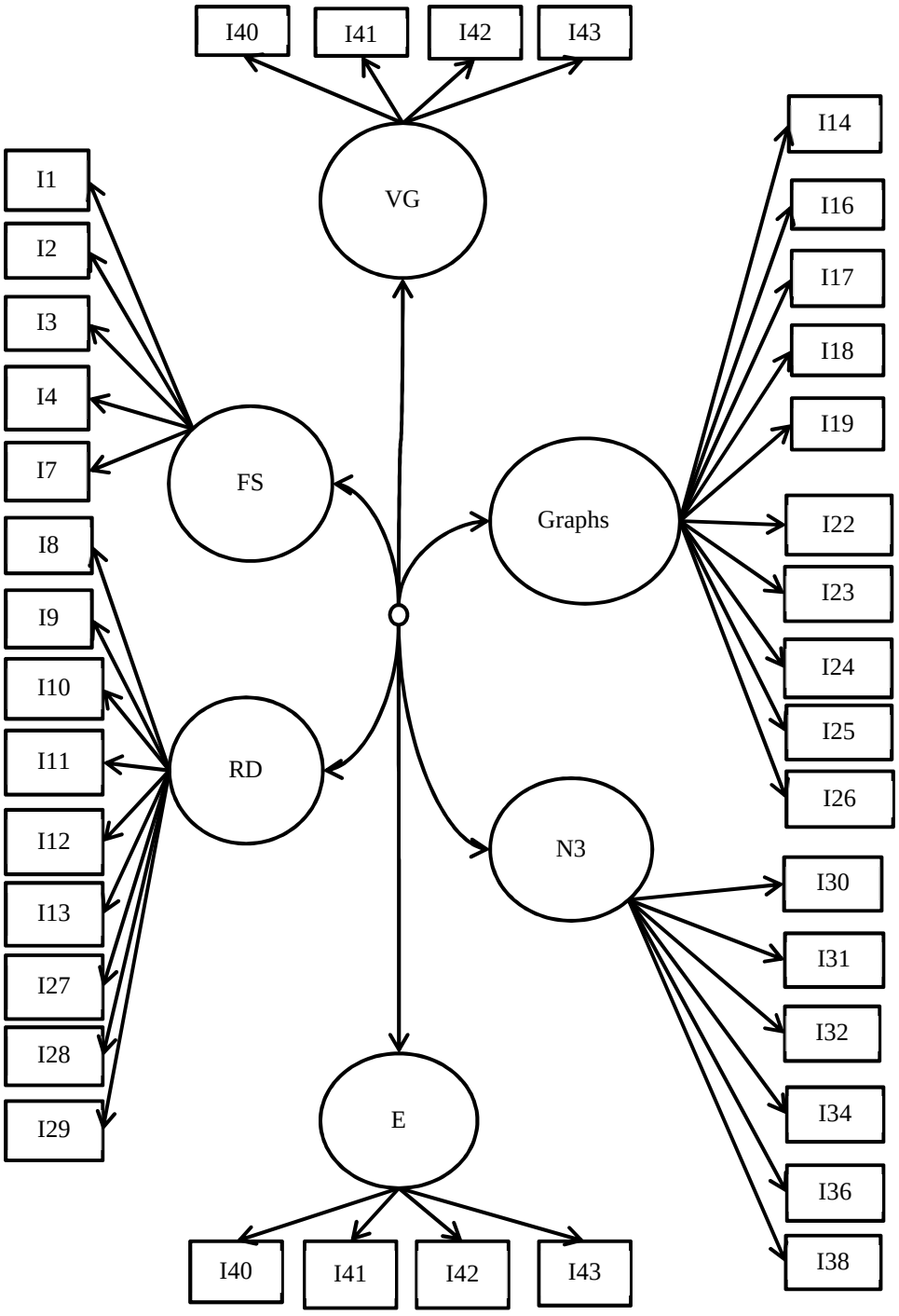


Figure B- 3. Path diagram for a confirmatory factor analysis.

The factor loadings of each item onto the latent factors are presented in

Table B- 8. Note that the model used a fixed factor variance strategy, so none of the regression weights were constrained to a value of one. Most loadings were in the 0.8 to 1.0 range, with the exception of the Velocity Graphs items that have been problematic in all analyses for their lack of variability. The covariances between the latent factors are presented in Table B-9 and are fairly strong. The Reverse Direction, Force Sled, and Force Graphs factors are related to each other with correlations near the 0.6 level. The FMCE testlets appear to be related to one another, in terms of student performance. Large modification indices produced by Mplus show that individual FMCE items are related to one another beyond their association through the related factors. These strong associations, through the factor structure and beyond it, indicate that the FMCE needs some uni-dimensional aspect to its modeling.

Table B- 8

Parameter estimates of a CFA model using combined pretest and posttest FMCE data

FS by	Weight	SE
FMCE1	0.974	0.006
FMCE2	0.960	0.008
FMCE3	0.862	0.010
FMCE4	0.957	0.006
FMCE7	0.832	0.011

RD by	Weight	SE
FMCE8	0.990	0.005
FMCE9	0.956	0.006
FMCE10	0.893	0.007
FMCE11	0.969	0.004
FMCE12	0.950	0.005
FMCE13	0.905	0.006
FMCE27	0.933	0.006
FMCE28	0.947	0.006
FMCE29	0.894	0.007

GRAPHS by	Weight	SE
FMCE14	0.984	0.004
FMCE16	0.958	0.005
FMCE17	0.965	0.005
FMCE18	0.967	0.006
FMCE19	0.928	0.007
FMCE22	0.945	0.005
FMCE23	0.931	0.005
FMCE24	0.920	0.005
FMCE25	0.875	0.008
FMCE26	0.917	0.006

N3 by	Weight	SE
FMCE30	0.913	0.007
FMCE31	0.857	0.010
FMCE32	0.980	0.004
FMCE34	0.987	0.004
FMCE36	0.911	0.012
FMCE38	0.921	0.012

VG by	Weight	SE
FMCE40	0.796	0.022
FMCE41	0.847	0.027
FMCE42	0.898	0.019
FMCE43	0.676	0.027

E by	Weight	SE
FMCE44	0.972	0.010
FMCE45	0.835	0.012
FMCE46	0.896	0.011
FMCE47	0.842	0.013

Table B- 9

Covariance parameter estimates (and standard errors) among latent factors in a CFA model using combined pretest and posttest data

	FS	RD	Graphs	N3	VG	E
FS						
RD	.784					
	(.010)					
Graphs	.820	.802				
	(.008)	(.008)				
N3	.579	.672	.617			
	(.016)	(.013)	(.014)			
VG	.414	.496	.527	.325		
	(.022)	(.019)	(.018)	(.022)		
E	.638	.666	.613	.542	.402	
	(.015)	(.014)	(.014)	(.017)	(.019)	

Bifactor modeling

There are a few models that can be used to explore instruments that are somewhere between uni-dimensional and multi-dimensional. The best fitting model in this case was a bifactor model, which uses one general factor which all items load upon and a series of specific latent factors to represent each unique component. *Figure B- 4*. Path diagram for a bifactor model of the FMCE shows the path diagram of the model used with a random half of the combined FMCE data. As before, the model was run in Mplus with a weighted least mean square estimation method. Note that the bifactor model does not include correlations among the latent variables. The model does include error variance parameters for each of the indicator variables, though they were not included in the path diagram to conserve space.

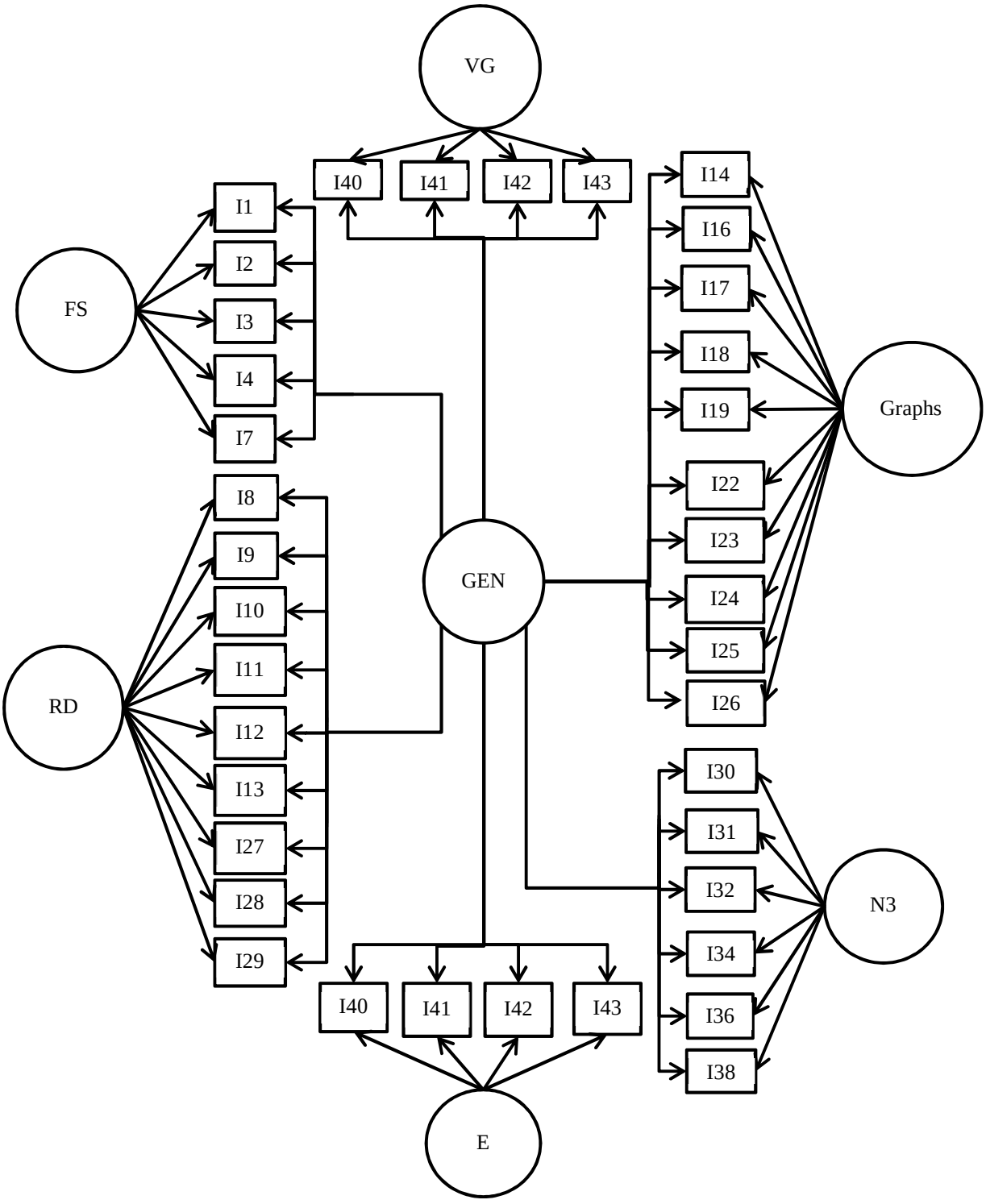


Figure B- 4. Path diagram for a bifactor model of the FMCE

The bifactor model produced a goodness-of-fit chi-square of 5240 with 627 degrees of freedom, still statistically significant. A chi-square difference test cannot be performed because the bifactor model is not nested within the CFA model. However, the RMSEA of .044 is below the guideline .05 value, and the CFI and TLI are .988 and .987, respectively. The bifactor model fits the data better than the CFA model. The parameter estimates from the bifactor model appear in Table B-10. Again, a fixed factor variance method does not require that any loadings be fixed at one. The loadings of the general factor are strong all items except those of the Velocity Graphs testlet. The other loadings in Table B-10 represent the effect of each local factor on each item after controlling for the general factor. These are noticeably smaller than the loadings for the general factor, but are generally non-zero.

Table B- 10

Parameter estimates from a bifactor model of combined pretest and posttest FMCE data

GEN by	Weight	SE
FMCE1	0.845	0.010
FMCE2	0.856	0.010
FMCE3	0.737	0.013
FMCE4	0.827	0.011
FMCE7	0.709	0.014
FMCE8	0.871	0.011
FMCE9	0.841	0.012
FMCE10	0.768	0.013
FMCE11	0.831	0.011
FMCE12	0.827	0.011
FMCE13	0.756	0.012
FMCE14	0.978	0.004
FMCE16	0.951	0.005
FMCE17	0.959	0.005
FMCE18	0.954	0.005
FMCE19	0.918	0.007
FMCE22	0.806	0.011
FMCE23	0.821	0.010
FMCE24	0.790	0.011
FMCE25	0.811	0.011
FMCE26	0.780	0.012
FMCE27	0.859	0.009
FMCE28	0.874	0.009
FMCE29	0.811	0.010
FMCE30	0.614	0.016
FMCE31	0.588	0.016
FMCE32	0.678	0.015
FMCE34	0.692	0.014
FMCE36	0.666	0.017
FMCE38	0.676	0.017

GEN by	Weight	SE
FMCE40	0.383	0.023
FMCE41	0.491	0.019
FMCE42	0.488	0.018
FMCE43	0.338	0.025
FMCE44	0.708	0.014
FMCE45	0.598	0.016
FMCE46	0.633	0.016
FMCE47	0.589	0.017

FS by	Weight	SE
FMCE1	0.494	0.017
FMCE2	0.331	0.017
FMCE3	0.505	0.019
FMCE4	0.491	0.016
FMCE7	0.504	0.020

RD by	Weight	SE
FMCE8	0.464	0.017
FMCE9	0.455	0.018
FMCE10	0.484	0.017
FMCE11	0.509	0.016
FMCE12	0.476	0.017
FMCE13	0.553	0.017
FMCE27	0.324	0.017
FMCE28	0.321	0.018
FMCE29	0.353	0.018

GRAPHS by	Weight	SE
FMCE14	-0.101	0.020
FMCE16	-0.055	0.018
FMCE17	-0.086	0.021
FMCE18	0.002	0.019
FMCE19	-0.031	0.020
FMCE22	0.555	0.016
FMCE23	0.492	0.016
FMCE24	0.545	0.016
FMCE25	0.382	0.017
FMCE26	0.554	0.016

N3 by	Weight	SE
FMCE30	0.703	0.014
FMCE31	0.646	0.015
FMCE32	0.714	0.014
FMCE34	0.701	0.014
FMCE36	0.571	0.018
FMCE38	0.567	0.018

VG by	Weight	SE
FMCE40	0.898	0.023
FMCE41	0.422	0.028
FMCE42	0.695	0.022
FMCE43	0.676	0.028

ENERGY by	Weight	SE
FMCE44	0.585	0.016
FMCE45	0.605	0.018
FMCE46	0.663	0.016
FMCE47	0.645	0.017

The most interesting feature of the bifactor model is the regression weights for the Force Graphs items which are not statistically significantly different from zero. The general factor has

completely taken over these items and they are no longer associated at all with the Graphs factor. In a sense, this solution asserts that the FMCE is centered on the Force Graphs items, which are directly related to the core concept of the test. The other items are determined by an understanding of that core concept but also influenced by other latent constructs. This result makes substantive sense because the Force Graphs testlet represents the core physics concept *and* the graphical format used for two other FMCE testlets. At the same time, it is surprising that Force Graphs and not Reverse Direction items that were most associated with a general factor, after seeing how the RD items were so strongly related to the rest of the instrument in the previous factor analyses.

Seeing this result, I ran a final model where Items 14 through 19 no longer loaded on any latent factor, only to the general factor. This model had a chi-square of 5141 with 632 degrees of freedom, statistically significantly better fitting than the previous bifactor model. The fit statistics show the model is better fitting: an RMSEA of .043, a CFI of .989 and a TLI of .987. Based on this evidence, I make the claim that this modified bifactor model is the best way to describe the FMCE in a continuous latent variable framework. The model, though it is the best fitting, is not easy to use in classroom or research settings. The scores generated by a bifactor model must be interpreted very carefully, where the specific factor scores are really the residuals of the scores after considering the general factor score (DeMars, 2013).

Appendix C: Sample Mplus mLTA Output

This appendix contains the output from one mLTA model, the six-class solution of the Force Sled testlet. It is annotated using text boxes, explaining the various parameters and results.

```
Mplus VERSION 7.3 (Linux)
MUTHEN & MUTHEN
01/13/2016   8:03 PM

INPUT INSTRUCTIONS

  TITLE:      Trial of an LTA for the force sled cluster

  DATA:       File = All data.dat;

  VARIABLE:   Names = ID Group PRE1-PRE47 POST1-POST47;
              Usevariable = Group PRE44-PRE47 POST44-POST47;
              Nominal = PRE44-PRE47 POST44-POST47;
              Classes = g(3) PRE(7) POST(7);
              Missing = All(-99);
           Knownclass = g(Group = 0 Group = 1 Group = 2);

  ANALYSIS:   Type = mixture;
              Starts = 200 20;
              Miterations = 1000;
              Processors = 24;

  MODEL:      %overall%
              POST PRE on g;

  OUTPUT:     TECH15;

  MODEL g:
              %g#1%
              POST on PRE;
              %g#2%
              POST on PRE;
              %g#3%
              POST on PRE;

  MODEL PRE:

  %PRE#1%
  [PRE44#1 PRE44#2 PRE44#3 PRE44#4] (1-4);
  [PRE45#1 PRE45#2 PRE45#3 PRE45#4] (5-8);
  [PRE46#1 PRE46#2 PRE46#3 PRE46#4] (9-12);
  [PRE47#1 PRE47#2 PRE47#3 PRE47#4] (13-16);

  %PRE#2%
  [PRE44#1 PRE44#2 PRE44#3 PRE44#4] (17-20);
```

All Mplus output includes the syntax that was used as input to define the model

This statement asks for pretest and posttest class membership to be predicted by group membership.

These statements ask for separate transition parameters for each treatment group.

These statements set measurement parameters equal across pretest and posttest

```
[PRE45#1 PRE45#2 PRE45#3 PRE45#4] (21-24);
[PRE46#1 PRE46#2 PRE46#3 PRE46#4] (25-28);
[PRE47#1 PRE47#2 PRE47#3 PRE47#4] (29-32);

%PRE#3%
[PRE44#1 PRE44#2 PRE44#3 PRE44#4] (33-36);
[PRE45#1 PRE45#2 PRE45#3 PRE45#4] (37-40);
[PRE46#1 PRE46#2 PRE46#3 PRE46#4] (41-44);
[PRE47#1 PRE47#2 PRE47#3 PRE47#4] (45-48);

%PRE#4%
[PRE44#1 PRE44#2 PRE44#3 PRE44#4] (49-52);
[PRE45#1 PRE45#2 PRE45#3 PRE45#4] (53-56);
[PRE46#1 PRE46#2 PRE46#3 PRE46#4] (57-60);
[PRE47#1 PRE47#2 PRE47#3 PRE47#4] (61-64);

%PRE#5%
[PRE44#1 PRE44#2 PRE44#3 PRE44#4] (65-68);
[PRE45#1 PRE45#2 PRE45#3 PRE45#4] (69-72);
[PRE46#1 PRE46#2 PRE46#3 PRE46#4] (73-76);
[PRE47#1 PRE47#2 PRE47#3 PRE47#4] (77-80);

%PRE#6%
[PRE44#1 PRE44#2 PRE44#3 PRE44#4] (81-84);
[PRE45#1 PRE45#2 PRE45#3 PRE45#4] (85-88);
[PRE46#1 PRE46#2 PRE46#3 PRE46#4] (89-92);
[PRE47#1 PRE47#2 PRE47#3 PRE47#4] (93-96);

%PRE#7%
[PRE44#1 PRE44#2 PRE44#3 PRE44#4] (97-100);
[PRE45#1 PRE45#2 PRE45#3 PRE45#4] (101-104);
[PRE46#1 PRE46#2 PRE46#3 PRE46#4] (105-108);
[PRE47#1 PRE47#2 PRE47#3 PRE47#4] (109-112);

MODEL POST:

%POST#1%
[POST44#1 POST44#2 POST44#3 POST44#4] (1-4);
[POST45#1 POST45#2 POST45#3 POST45#4] (5-8);
[POST46#1 POST46#2 POST46#3 POST46#4] (9-12);
[POST47#1 POST47#2 POST47#3 POST47#4] (13-16);

%POST#2%
[POST44#1 POST44#2 POST44#3 POST44#4] (17-20);
[POST45#1 POST45#2 POST45#3 POST45#4] (21-24);
[POST46#1 POST46#2 POST46#3 POST46#4] (25-28);
[POST47#1 POST47#2 POST47#3 POST47#4] (29-32);


%POST#3%
[POST44#1 POST44#2 POST44#3 POST44#4] (33-36);
[POST45#1 POST45#2 POST45#3 POST45#4] (37-40);
```

```
  [POST46#1 POST46#2 POST46#3 POST46#4] (41-44);
  [POST47#1 POST47#2 POST47#3 POST47#4] (45-48);

  %POST#4%
  [POST44#1 POST44#2 POST44#3 POST44#4] (49-52);
  [POST45#1 POST45#2 POST45#3 POST45#4] (53-56);
  [POST46#1 POST46#2 POST46#3 POST46#4] (57-60);
  [POST47#1 POST47#2 POST47#3 POST47#4] (61-64);

  %POST#5%
  [POST44#1 POST44#2 POST44#3 POST44#4] (65-68);
  [POST45#1 POST45#2 POST45#3 POST45#4] (69-72);
  [POST46#1 POST46#2 POST46#3 POST46#4] (73-76);
  [POST47#1 POST47#2 POST47#3 POST47#4] (77-80);

  %POST#6%
  [POST44#1 POST44#2 POST44#3 POST44#4] (81-84);
  [POST45#1 POST45#2 POST45#3 POST45#4] (85-88);
  [POST46#1 POST46#2 POST46#3 POST46#4] (89-92);
  [POST47#1 POST47#2 POST47#3 POST47#4] (93-96);

  %POST#7%
  [POST44#1 POST44#2 POST44#3 POST44#4] (97-100);
  [POST45#1 POST45#2 POST45#3 POST45#4] (101-104);
  [POST46#1 POST46#2 POST46#3 POST46#4] (105-108);
  [POST47#1 POST47#2 POST47#3 POST47#4] (109-112);
```

The posttest measurement parameters are held to equal the pretest parameters by the parenthetical statements.

The output begins here with a description of the data and the model.

```
Trial of an LTA for the force sled cluster

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        4516

Number of dependent variables                                    8
Number of independent variables                                  0
Number of continuous latent variables                            0
Number of categorical latent variables                           3

Observed dependent variables

  Unordered categorical (nominal)
   PRE44       PRE45       PRE46       PRE47       POST44      POST45
   POST46      POST47

Categorical latent variables
```

```
   G           PRE          POST

  Knownclass            G
```

This section specifies values used in the estimation, such as convergence criteria.

```
Estimator                                                      MLR
Information matrix                                        OBSERVED
Optimization Specifications for the Quasi-Newton Algorithm for
Continuous Outcomes
  Maximum number of iterations                                 100
  Convergence criterion                                0.100D-05
Optimization Specifications for the EM Algorithm
  Maximum number of iterations                                1000
  Convergence criteria
    Loglikelihood change                               0.100D-06
    Relative loglikelihood change                      0.100D-06
    Derivative                                         0.100D-05
Optimization Specifications for the M step of the EM Algorithm for
Categorical Latent variables
  Number of M step iterations                                    1
  M step convergence criterion                         0.100D-05
  Basis for M step termination                         ITERATION
Optimization Specifications for the M step of the EM Algorithm for
Censored, Binary or Ordered Categorical (Ordinal), Unordered
Categorical (Nominal) and Count Outcomes
  Number of M step iterations                                    1
  M step convergence criterion                         0.100D-05
  Basis for M step termination                         ITERATION
  Maximum value for logit thresholds                          15
  Minimum value for logit thresholds                         -15
  Minimum expected cell size for chi-square            0.100D-01
Maximum number of iterations for H1                         2000
Convergence criterion for H1                           0.100D-03
Optimization algorithm                                       EMA
Random Starts Specifications
  Number of initial stage random starts                       200
  Number of final stage optimizations                          20
  Number of initial stage iterations                           10
  Initial stage convergence criterion                  0.100D+01
  Random starts scale                                  0.500D+01
  Random seed for generating random starts                     0
Parameterization                                            LOGIT

Input data file(s)
  All data.dat
Input data format  FREE


SUMMARY OF DATA
```

```
     Number of missing data patterns             0
     Number of y missing data patterns           0
     Number of u missing data patterns           0


COVARIANCE COVERAGE OF DATA

Minimum covariance coverage value   0.100


UNIVARIATE PROPORTIONS AND COUNTS FOR CATEGORICAL VARIABLES

    PRE44
      Category 1    0.669     2693.000
      Category 2    0.257     1035.000
      Category 3    0.036      146.000
      Category 4    0.030      121.000
      Category 5    0.007       28.000
    PRE45
      Category 1    0.496     1869.000
      Category 2    0.404     1523.000
      Category 3    0.057      216.000
      Category 4    0.036      135.000
      Category 5    0.007       28.000
   …

   …

RANDOM STARTS RESULTS RANKED FROM THE BEST TO THE WORST LOGLIKELIHOOD
VALUES

Final stage loglikelihood values at local maxima, seeds, and initial
stage start numbers:

          -30408.011  645664           39
          -30408.011  695155           150
          -30408.011  456213           160
          -30408.011  570782           193
          -30408.011  314084           81
          -30408.011  164305           128
          -30408.012  370466           41
          -30408.107  195873           6
          -30408.107  957392           79
          -30408.156  100874           108
          -30408.156  347515           24
          -30462.132  85462            51
          -30486.647  297518           166
```

These values, abbreviated, give the number of observed responses for each option of each item.

This table provides the seed values for running specific replications and the loglikelihood for each converged replication.

```
          -30522.526  526324           178
          -30569.760  40340            188
          -30572.914  939709           112
          -30636.527  565819           65

3 perturbed starting value run(s) did not converge.
```

Of the 20 replications selected, only 17 converged on a solution.

```
THE BEST LOGLIKELIHOOD VALUE HAS BEEN REPLICATED.  RERUN WITH AT LEAST
TWICE THE
RANDOM STARTS TO CHECK THAT THE BEST LOGLIKELIHOOD IS STILL OBTAINED
AND REPLICATED.

     IN THE OPTIMIZATION, ONE OR MORE LOGIT SCALE PARAMETERS
APPROACHED AND WERE
     SET AT THE EXTREME VALUES.  EXTREME VALUES ARE -15.000 AND
15.000.
     THE FOLLOWING PARAMETERS WERE SET AT THESE VALUES:
     * MEAN OF POST44#1 FOR PATTERN 1 1 2
     * MEAN OF POST44#3 FOR PATTERN 1 1 2
     * MEAN OF POST45#1 FOR PATTERN 1 1 2
     * MEAN OF POST46#3 FOR PATTERN 1 1 2
     * MEAN OF POST46#2 FOR PATTERN 1 1 3
     * MEAN OF POST46#2 FOR PATTERN 1 1 5
     * MEAN OF POST46#4 FOR PATTERN 1 1 5
     * MEAN OF POST47#4 FOR PATTERN 1 1 5
       ...
     * MEAN OF POST46#4 FOR PATTERN 3 7 6
     * MEAN OF POST47#1 FOR PATTERN 3 7 6
     * MEAN OF POST47#3 FOR PATTERN 3 7 6
     * MEAN OF POST47#4 FOR PATTERN 3 7 6
     * MEAN OF PRE45#4 FOR PATTERN 3 7 7
     * MEAN OF PRE47#4 FOR PATTERN 3 7 7
     * MEAN OF POST45#4 FOR PATTERN 3 7 7
     * MEAN OF POST47#4 FOR PATTERN 3 7 7
```

Dozens of parameters were fixed at 15 or -15, defining classes by responses that are always or never selected.

```
ONE OR MORE MULTINOMIAL LOGIT PARAMETERS WERE FIXED TO AVOID
SINGULARITY OF THE INFORMATION MATRIX.  THE SINGULARITY IS MOST LIKELY
BECAUSE THE MODEL IS NOT IDENTIFIED, OR BECAUSE OF EMPTY CELLS IN THE
JOINT DISTRIBUTION OF THE CATEGORICAL LATENT VARIABLES AND ANY
INDEPENDENT VARIABLES.  THE FOLLOWING PARAMETERS WERE FIXED:
Parameter 51, MODEL G: %G#1%: POST#1 ON PRE#3
     Parameter 63, MODEL G: %G#1%: POST#1 ON PRE#5
```

```
     Parameter 70, MODEL G: %G#1%: POST#2 ON PRE#6
     Parameter 72, MODEL G: %G#1%: POST#4 ON PRE#6
     Parameter 87, MODEL G: %G#2%: POST#1 ON PRE#3
     Parameter 105, MODEL G: %G#2%: POST#1 ON PRE#6
     Parameter 106, MODEL G: %G#2%: POST#2 ON PRE#6
     Parameter 107, MODEL G: %G#2%: POST#3 ON PRE#6
     Parameter 110, MODEL G: %G#2%: POST#6 ON PRE#6
     Parameter 143, MODEL G: %G#3%: POST#3 ON PRE#6
     Parameter 144, MODEL G: %G#3%: POST#4 ON PRE#6
```

These parameters were fixed to allow the information matrix to be inverted. They are all transition parameters.

```
THE MODEL ESTIMATION TERMINATED NORMALLY

MODEL FIT INFORMATION

Number of Free Parameters                       258

Loglikelihood

          H0 Value                       -30408.011
          H0 Scaling Correction Factor        1.0309
            for MLR

Information Criteria

          Akaike (AIC)                     61332.023
          Bayesian (BIC)                   62987.191
          Sample-Size Adjusted BIC         62167.368
           (n* = (n + 2) / 24)
```

BIC used to decide on 7 class solution.

```
FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASS PATTERNS
BASED ON THE ESTIMATED MODEL
```

Classes labeled by treatment group, pretest class, and posttest class.

```
  Latent Class
    Pattern

    1  1  1          65.05898          0.01441
    1  1  2          42.53631          0.00942
    1  1  3          17.74468          0.00393
    1  1  4          13.98653          0.00310
    1  1  5          48.46413          0.01073
    1  1  6          13.85505          0.00307
    1  1  7          34.46941          0.00763
    1  2  1           6.40273          0.00142
    1  2  2          53.71462          0.01189
    1  2  3          13.18552          0.00292
```

1	2	4	15.88178	0.00352
1	2	5	30.30384	0.00671
1	2	6	16.65400	0.00369
1	2	7	36.61124	0.00811
1	3	1	0.00000	0.00000
1	3	2	22.80756	0.00505
1	3	3	25.07369	0.00555
1	3	4	31.27770	0.00693
1	3	5	31.24996	0.00692
1	3	6	11.69155	0.00259
1	3	7	41.06681	0.00909
1	4	1	15.55455	0.00344
1	4	2	4.87534	0.00108
1	4	3	14.89088	0.00330
1	4	4	46.54409	0.01031
1	4	5	15.53546	0.00344
1	4	6	9.79478	0.00217
1	4	7	69.00467	0.01528
1	5	1	0.00000	0.00000
1	5	2	0.28181	0.00006
1	5	3	7.72537	0.00171
1	5	4	6.22760	0.00138
1	5	5	17.22363	0.00381
1	5	6	3.82974	0.00085
1	5	7	18.60503	0.00412
1	6	1	3.05109	0.00068
1	6	2	5.52257	0.00122
1	6	3	0.63398	0.00014
1	6	4	0.90535	0.00020
1	6	5	8.12410	0.00180
1	6	6	12.35564	0.00274
1	6	7	17.01801	0.00377
1	7	1	2.65015	0.00059
1	7	2	0.00000	0.00000
1	7	3	0.64071	0.00014
1	7	4	0.00000	0.00000
1	7	5	7.84209	0.00174
1	7	6	0.89704	0.00020
1	7	7	100.23022	0.02219
2	1	1	96.44029	0.02136
2	1	2	20.87010	0.00462
2	1	3	17.23929	0.00382
2	1	4	11.01649	0.00244
2	1	5	48.60685	0.01076
2	1	6	13.46868	0.00298
2	1	7	82.76057	0.01833
2	2	1	11.75745	0.00260
2	2	2	50.83776	0.01126
2	2	3	35.82572	0.00793
2	2	4	13.60497	0.00301
2	2	5	32.16462	0.00712
2	2	6	15.32240	0.00339

The proportion of students in each treatment group, pretest class and posttest class, as predicted by the model solution.

2	2	7	94.32120	0.02089
2	3	1	0.00000	0.00000
2	3	2	23.67425	0.00524
2	3	3	36.42509	0.00807
2	3	4	22.15456	0.00491
2	3	5	39.82610	0.00882
2	3	6	11.75293	0.00260
2	3	7	133.44071	0.02955
2	4	1	13.66938	0.00303
2	4	2	22.18151	0.00491
2	4	3	21.42247	0.00474
2	4	4	65.39199	0.01448
2	4	5	32.93288	0.00729
2	4	6	26.21444	0.00580
2	4	7	166.47083	0.03686
2	5	1	12.38543	0.00274
2	5	2	7.00536	0.00155
2	5	3	4.86667	0.00108
2	5	4	10.38916	0.00230
2	5	5	27.97956	0.00620
2	5	6	8.09510	0.00179
2	5	7	35.84840	0.00794
2	6	1	0.00000	0.00000
2	6	2	7.24096	0.00160
2	6	3	3.96596	0.00088
2	6	4	1.16830	0.00026
2	6	5	8.10897	0.00180
2	6	6	0.00000	0.00000
2	6	7	27.31808	0.00605
2	7	1	6.33388	0.00140
2	7	2	0.00000	0.00000
2	7	3	0.00000	0.00000
2	7	4	0.88511	0.00020
2	7	5	11.54222	0.00256
2	7	6	5.31860	0.00118
2	7	7	202.75472	0.04490
3	1	1	97.13700	0.02151
3	1	2	11.30484	0.00250
3	1	3	26.34996	0.00583
3	1	4	4.85362	0.00107
3	1	5	13.67652	0.00303
3	1	6	21.30088	0.00472
3	1	7	124.89330	0.02766
3	2	1	34.63911	0.00767
3	2	2	39.15438	0.00867
3	2	3	8.44166	0.00187
3	2	4	8.84668	0.00196
3	2	5	38.34419	0.00849
3	2	6	19.91469	0.00441
3	2	7	119.97478	0.02657
3	3	1	38.72407	0.00857
3	3	2	28.41209	0.00629

```
    3  3  3         18.47874          0.00409
    3  3  4         30.02489          0.00665
    3  3  5         47.53488          0.01053
    3  3  6         28.69884          0.00635
    3  3  7        164.93380          0.03652
    3  4  1         25.26626          0.00559
    3  4  2          5.54768          0.00123
    3  4  3         30.47917          0.00675
    3  4  4         48.46877          0.01073
    3  4  5         43.63197          0.00966
    3  4  6         16.65600          0.00369
    3  4  7        198.43611          0.04394
    3  5  1         16.86804          0.00374
    3  5  2          8.36014          0.00185
    3  5  3          2.63762          0.00058
    3  5  4         23.26694          0.00515
    3  5  5         64.30136          0.01424
    3  5  6          8.87559          0.00197
    3  5  7        111.50540          0.02469
    3  6  1          9.00187          0.00199
    3  6  2          8.17164          0.00181
    3  6  3          0.00000          0.00000
    3  6  4          0.00000          0.00000
    3  6  5         12.11470          0.00268
    3  6  6          5.92671          0.00131
    3  6  7         72.78584          0.01612
    3  7  1         17.60071          0.00390
    3  7  2          5.76746          0.00128
    3  7  3          7.23199          0.00160
    3  7  4          8.14867          0.00180
    3  7  5         15.37586          0.00340
    3  7  6          5.70330          0.00126
    3  7  7        315.23129          0.06980


FINAL CLASS COUNTS AND PROPORTIONS FOR EACH LATENT CLASS VARIABLE
BASED ON THE ESTIMATED MODEL

  Latent Class
    Variable    Class

    G               1       961.99994          0.21302
                    2      1541.00024          0.34123
                    3      2013.00000          0.44575
    PRE             1       826.03357          0.18291
                    2       695.90332          0.15410
                    3       787.24817          0.17432
                    4       892.96924          0.19773
                    5       396.27792          0.08775
                    6       203.41376          0.04504
                    7       714.15405          0.15814
    POST            1       472.54099          0.10464
```

```
          2        368.26633          0.08155
          3        293.25919          0.06494
          4        363.04321          0.08039
          5        594.88385          0.13173
          6        256.32596          0.05676
          7       2167.68042          0.48000
```

These classification tables provide some information about student membership. The transition probability tables appear at the end of the output file.

```
LATENT TRANSITION PROBABILITIES BASED ON THE ESTIMATED MODEL

  G Classes (Rows) by PRE Classes (Columns)

        1        2        3        4        5        6        7

   1  0.245    0.180    0.170    0.183    0.056    0.049    0.117
   2  0.188    0.165    0.173    0.226    0.069    0.031    0.147
   3  0.149    0.134    0.177    0.183    0.117    0.054    0.186


  PRE Classes (Rows) by POST Classes (Columns)

        1        2        3        4        5        6        7

   1  0.313    0.090    0.074    0.036    0.134    0.059    0.293
   2  0.076    0.207    0.083    0.055    0.145    0.075    0.361
   3  0.049    0.095    0.102    0.106    0.151    0.066    0.431
   4  0.061    0.037    0.075    0.180    0.103    0.059    0.486
   5  0.074    0.039    0.038    0.101    0.276    0.052    0.419
   6  0.059    0.103    0.023    0.010    0.139    0.090    0.576
   7  0.037    0.008    0.011    0.013    0.049    0.017    0.866


FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASS PATTERNS
BASED ON ESTIMATED POSTERIOR PROBABILITIES

  Latent Class
    Pattern

    1  1  1         65.05899          0.01441
    1  1  2         42.53631          0.00942
    1  1  3         17.74468          0.00393
    1  1  4         13.98653          0.00310
```

...

```
    3   6   5          12.11470          0.00268
    3   6   6           5.92671          0.00131
    3   6   7          72.78585          0.01612
    3   7   1          17.60071          0.00390
    3   7   2           5.76746          0.00128
    3   7   3           7.23199          0.00160
    3   7   4           8.14867          0.00180
    3   7   5          15.37585          0.00340
    3   7   6           5.70330          0.00126
    3   7   7         315.23129          0.06980
```

These proportions, abbreviated, are based on posterior probabilities of students in the sample.

```
FINAL CLASS COUNTS AND PROPORTIONS FOR EACH LATENT CLASS VARIABLE
BASED ON ESTIMATED POSTERIOR PROBABILITIES

  Latent Class
    Variable    Class

    G               1        961.99994          0.21302
                    2       1541.00024          0.34123
                    3       2013.00000          0.44575
    PRE             1        826.03357          0.18291
                    2        695.90332          0.15410
                    3        787.24817          0.17432
                    4        892.96924          0.19773
                    5        396.27792          0.08775
                    6        203.41376          0.04504
                    7        714.15405          0.15814
    POST            1        472.54102          0.10464
                    2        368.26633          0.08155
                    3        293.25919          0.06494
                    4        363.04321          0.08039
                    5        594.88385          0.13173
                    6        256.32596          0.05676
                    7       2167.68042          0.48000


FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASS PATTERNS
BASED ON THEIR MOST LIKELY LATENT CLASS PATTERN

Class Counts and Proportions

  Latent Class
    Pattern

    1   1   1                89          0.01971
    1   1   2                33          0.00731
    1   1   3                 9          0.00199
    1   1   4                 7          0.00155
    1   1   5                48          0.01063
```

...

These proportions, abbreviated, are based on most the modal posterior probability for each student in the sample.

```
    3  7  1              9          0.00199
    3  7  2              3          0.00066
    3  7  3              4          0.00089
    3  7  4              5          0.00111
    3  7  5              8          0.00177
    3  7  6              3          0.00066
    3  7  7            459          0.10164


FINAL CLASS COUNTS AND PROPORTIONS FOR EACH LATENT CLASS VARIABLE
BASED ON THEIR MOST LIKELY LATENT CLASS PATTERN

  Latent Class
    Variable    Class

    G               1            962          0.21302
                    2           1541          0.34123
                    3           2013          0.44575
    PRE             1            715          0.15833
                    2            650          0.14393
                    3            808          0.17892
                    4            889          0.19686
                    5            381          0.08437
                    6            204          0.04517
                    7            869          0.19243
    POST            1            379          0.08392
                    2            324          0.07174
                    3            253          0.05602
                    4            301          0.06665
                    5            477          0.10562
                    6            220          0.04872
                    7           2562          0.56732


CLASSIFICATION QUALITY

     Entropy                         0.817

Average Latent Class Probabilities for Most Likely Latent Class
Pattern (Row)
by Latent Class Pattern (Column)
```

The entropy value, above the 0.8 guideline indicates small amounts of cross classification.

This section gives values for each of the measurement parameters. This first set is for NOT BOTH students who transitioned from Class 1 to Class 1.

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Latent Class Pattern 1 1 1				
Means				
PRE44#1	2.771	0.318	8.703	0.000
PRE44#2	2.457	0.287	8.572	0.000
PRE44#3	1.954	0.262	7.461	0.000
PRE44#4	1.211	0.285	4.257	0.000
PRE45#1	2.284	0.247	9.231	0.000
PRE45#2	1.780	0.332	5.365	0.000
PRE45#3	1.973	0.203	9.705	0.000
PRE45#4	1.261	0.209	6.043	0.000
PRE46#1	1.763	0.188	9.358	0.000
PRE46#2	1.203	0.197	6.107	0.000
PRE46#3	1.305	0.229	5.688	0.000
PRE46#4	0.508	0.296	1.716	0.086
PRE47#1	1.589	0.204	7.782	0.000
PRE47#2	1.126	0.242	4.658	0.000
PRE47#3	1.654	0.194	8.531	0.000
PRE47#4	0.962	0.196	4.908	0.000
POST44#1	2.771	0.318	8.703	0.000
POST44#2	2.457	0.287	8.572	0.000
POST44#3	1.954	0.262	7.461	0.000
POST44#4	1.211	0.285	4.257	0.000
POST45#1	2.284	0.247	9.231	0.000
POST45#2	1.780	0.332	5.365	0.000
POST45#3	1.973	0.203	9.705	0.000
POST45#4	1.261	0.209	6.043	0.000
POST46#1	1.763	0.188	9.358	0.000
POST46#2	1.203	0.197	6.107	0.000
POST46#3	1.305	0.229	5.688	0.000
POST46#4	0.508	0.296	1.716	0.086
POST47#1	1.589	0.204	7.782	0.000
POST47#2	1.126	0.242	4.658	0.000
POST47#3	1.654	0.194	8.531	0.000
POST47#4	0.962	0.196	4.908	0.000
...				
Latent Class Pattern 2 1 1				
Means				
PRE44#1	2.771	0.318	8.703	0.000
PRE44#2	2.457	0.287	8.572	0.000
PRE44#3	1.954	0.262	7.461	0.000
PRE44#4	1.211	0.285	4.257	0.000
PRE45#1	2.284	0.247	9.231	0.000

The PRE and POST parameters are equal because both are the same class and were constrained above.

This section has the parameters for the BOTH students transitioning from Class 1 to Class 1.

PRE45#2	1.780	0.332	5.365	0.000
PRE45#3	1.973	0.203	9.705	0.000
PRE45#4	1.261	0.209	6.043	0.000
PRE46#1	1.763	0.188	9.358	0.000
PRE46#2	1.203	0.197	6.107	0.000
PRE46#3	1.305	0.229	5.688	0.000
PRE46#4	0.508	0.296	1.716	0.086
PRE47#1	1.589	0.204	7.782	0.000
PRE47#2	1.126	0.242	4.658	0.000
PRE47#3	1.654	0.194	8.531	0.000
PRE47#4	0.962	0.196	4.908	0.000
POST44#1	2.771	0.318	8.703	0.000
POST44#2	2.457	0.287	8.572	0.000
POST44#3	1.954	0.262	7.461	0.000
POST44#4	1.211	0.285	4.257	0.000
POST45#1	2.284	0.247	9.231	0.000
POST45#2	1.780	0.332	5.365	0.000
POST45#3	1.973	0.203	9.705	0.000
POST45#4	1.261	0.209	6.043	0.000
POST46#1	1.763	0.188	9.358	0.000
POST46#2	1.203	0.197	6.107	0.000
POST46#3	1.305	0.229	5.688	0.000
POST46#4	0.508	0.296	1.716	0.086
POST47#1	1.589	0.204	7.782	0.000
POST47#2	1.126	0.242	4.658	0.000
POST47#3	1.654	0.194	8.531	0.000
POST47#4	0.962	0.196	4.908	0.000

The parameters here are equal to those above because they have been constrained across treatment groups.

Latent Class Pattern 2 1 2

Means

PRE44#1	2.771	0.318	8.703	0.000
PRE44#2	2.457	0.287	8.572	0.000
PRE44#3	1.954	0.262	7.461	0.000
PRE44#4	1.211	0.285	4.257	0.000
PRE45#1	2.284	0.247	9.231	0.000
PRE45#2	1.780	0.332	5.365	0.000
PRE45#3	1.973	0.203	9.705	0.000
PRE45#4	1.261	0.209	6.043	0.000
PRE46#1	1.763	0.188	9.358	0.000
PRE46#2	1.203	0.197	6.107	0.000
PRE46#3	1.305	0.229	5.688	0.000
PRE46#4	0.508	0.296	1.716	0.086
PRE47#1	1.589	0.204	7.782	0.000
PRE47#2	1.126	0.242	4.658	0.000
PRE47#3	1.654	0.194	8.531	0.000
PRE47#4	0.962	0.196	4.908	0.000
POST44#1	15.000	0.000	999.000	999.000
POST44#2	11.260	0.546	20.610	0.000
POST44#3	-15.000	0.000	999.000	999.000
POST44#4	9.414	1.009	9.333	0.000
POST45#1	15.000	0.000	999.000	999.000

```
    POST45#2          11.360      0.880     12.902      0.000
    POST45#3           9.301      2.429      3.829      0.000
    POST45#4          10.228      0.631     16.216      0.000
    POST46#1          11.350      0.519     21.855      0.000
    POST46#2          12.265      0.318     38.559      0.000
    POST46#3          15.000      0.000    999.000    999.000
    POST46#4          11.936      0.543     21.967      0.000
    POST47#1           2.279      0.530      4.298      0.000
    POST47#2           2.121      0.577      3.677      0.000
    POST47#3           4.461      0.454      9.815      0.000
    POST47#4          -0.332      1.451     -0.229      0.819
```

...

```
Latent Class Pattern 3 1 2

 Means
    PRE44#1            2.771      0.318      8.703      0.000
    PRE44#2            2.457      0.287      8.572      0.000
    PRE44#3            1.954      0.262      7.461      0.000
    PRE44#4            1.211      0.285      4.257      0.000
    PRE45#1            2.284      0.247      9.231      0.000
    PRE45#2            1.780      0.332      5.365      0.000
    PRE45#3            1.973      0.203      9.705      0.000
    PRE45#4            1.261      0.209      6.043      0.000
    PRE46#1            1.763      0.188      9.358      0.000
    PRE46#2            1.203      0.197      6.107      0.000
    PRE46#3            1.305      0.229      5.688      0.000
    PRE46#4            0.508      0.296      1.716      0.086
    PRE47#1            1.589      0.204      7.782      0.000
    PRE47#2            1.126      0.242      4.658      0.000
    PRE47#3            1.654      0.194      8.531      0.000
    PRE47#4            0.962      0.196      4.908      0.000
    POST44#1          15.000      0.000    999.000    999.000
    POST44#2          11.260      0.546     20.610      0.000
    POST44#3         -15.000      0.000    999.000    999.000
    POST44#4           9.414      1.009      9.333      0.000
    POST45#1          15.000      0.000    999.000    999.000
    POST45#2          11.360      0.880     12.902      0.000
    POST45#3           9.301      2.429      3.829      0.000
    POST45#4          10.228      0.631     16.216      0.000
    POST46#1          11.350      0.519     21.855      0.000
    POST46#2          12.265      0.318     38.559      0.000
    POST46#3          15.000      0.000    999.000    999.000
    POST46#4          11.936      0.543     21.967      0.000
    POST47#1           2.279      0.530      4.298      0.000
    POST47#2           2.121      0.577      3.677      0.000
    POST47#3           4.461      0.454      9.815      0.000
    POST47#4          -0.332      1.451     -0.229      0.819
```

...

```
Latent Class Pattern 3 7 7

 Means
    PRE44#1            3.533      6.403      0.552      0.581
    PRE44#2            9.271      5.902      1.571      0.116
    PRE44#3            2.989      5.932      0.504      0.614
    PRE44#4            2.016      6.112      0.330      0.741
    PRE45#1            2.201      3.730      0.590      0.555
    PRE45#2            8.678      2.646      3.280      0.001
    PRE45#3            2.222      2.789      0.796      0.426
    PRE45#4          -15.000      0.000    999.000    999.000
    PRE46#1            7.130      1.030      6.925      0.000
    PRE46#2            0.653      1.833      0.356      0.722
    PRE46#3            2.181      1.104      1.976      0.048
    PRE46#4            0.992      1.330      0.746      0.456
    PRE47#1            7.195      0.869      8.278      0.000
    PRE47#2            3.708      0.886      4.184      0.000
    PRE47#3            2.162      0.959      2.255      0.024
    PRE47#4          -15.000      0.000    999.000    999.000
    POST44#1           3.533      6.403      0.552      0.581
    POST44#2           9.271      5.902      1.571      0.116
    POST44#3           2.989      5.932      0.504      0.614
    POST44#4           2.016      6.112      0.330      0.741
    POST45#1           2.201      3.730      0.590      0.555
    POST45#2           8.678      2.646      3.280      0.001
    POST45#3           2.222      2.789      0.796      0.426
    POST45#4         -15.000      0.000    999.000    999.000
    POST46#1           7.130      1.030      6.925      0.000
    POST46#2           0.653      1.833      0.356      0.722
    POST46#3           2.181      1.104      1.976      0.048
    POST46#4           0.992      1.330      0.746      0.456
    POST47#1           7.195      0.869      8.278      0.000
    POST47#2           3.708      0.886      4.184      0.000
    POST47#3           2.162      0.959      2.255      0.024
    POST47#4         -15.000      0.000    999.000    999.000

Categorical Latent Variables

 POST#1   ON
    G#1               -0.747      0.971     -0.770      0.441
    G#2               -0.581      0.690     -0.841      0.400

 POST#2   ON
    G#1              -24.279      1.007    -24.112      0.000
    G#2              -24.417      0.851    -28.676      0.000

 POST#3   ON
    G#1               -1.278      2.834     -0.451      0.652
```

These parameters predict posttest class membership by treatment group, the effect beyond the class means and transitions.

These parameters are in reference to the TUTORIAL group.

```
   G#2              -31.336      1.084    -28.918      0.000

POST#4   ON
   G#1              -24.503      1.616    -15.165      0.000
   G#2               -1.779      1.421     -1.252      0.211

POST#5   ON
   G#1                0.473      0.596      0.793      0.428
   G#2                0.155      0.546      0.283      0.777

POST#6   ON
   G#1               -0.704      1.557     -0.452      0.651
   G#2                0.371      0.919      0.404      0.686

PRE#1    ON
   G#1                0.968      0.187      5.178      0.000
   G#2                0.472      0.161      2.935      0.003

PRE#2    ON
   G#1                0.762      0.168      4.531      0.000
   G#2                0.444      0.143      3.096      0.002

PRE#3    ON
   G#1                0.424      0.168      2.516      0.012
   G#2                0.214      0.142      1.505      0.132

PRE#4    ON
   G#1                0.468      0.158      2.967      0.003
   G#2                0.446      0.126      3.544      0.000

PRE#5    ON
   G#1               -0.270      0.225     -1.199      0.231
   G#2               -0.291      0.175     -1.662      0.096

PRE#6    ON
   G#1                0.387      0.296      1.310      0.190
   G#2               -0.312      0.267     -1.169      0.243

Means
   G#1               -0.738      0.039    -18.838      0.000
   G#2               -0.267      0.034     -7.894      0.000
   PRE#1             -0.225      0.184     -1.223      0.221
   PRE#2             -0.331      0.160     -2.070      0.038
   PRE#3             -0.050      0.252     -0.198      0.843
   PRE#4             -0.018      0.110     -0.160      0.873
   PRE#5             -0.464      0.165     -2.815      0.005
   PRE#6             -1.245      0.280     -4.451      0.000
```

These parameters predict pretest class membership by treatment group.

The mean membership parameters below are in reference to the TUTORIAL group or to Class 7.

```
    POST#1            -2.885      0.459     -6.283      0.000
    POST#2            -4.001      0.648     -6.171      0.000
    POST#3            -3.775      0.776     -4.866      0.000
    POST#4            -3.655      0.574     -6.368      0.000
    POST#5            -3.021      0.413     -7.317      0.000
    POST#6            -4.012      0.743     -5.400      0.000
```

This heading is misleading, the parameters are for all students in Group 1 (TUTORIAL).

The transition parameters below describe the impact of pretest membership on posttest membership.

```
Latent Class Pattern 1 1 1

 POST#1   ON
    PRE#1            4.268      0.940      4.540      0.000
    PRE#2            1.889      1.220      1.549      0.121
    PRE#3          -23.228      0.000    999.000    999.000
    PRE#4            2.143      0.984      2.178      0.029
    PRE#5          -22.468      0.000    999.000    999.000
    PRE#6            1.914      1.555      1.231      0.218

 POST#2   ON
    PRE#1           28.491      0.888     32.085      0.000
    PRE#2           28.664      0.814     35.196      0.000
    PRE#3           27.692      0.869     31.861      0.000
    PRE#4           25.630      1.085     23.625      0.000
    PRE#5           24.091      9.957      2.420      0.016
    PRE#6           27.155      0.000    999.000    999.000

 POST#3   ON
    PRE#1            4.389      2.811      1.561      0.118
    PRE#2            4.031      2.836      1.422      0.155
    PRE#3            4.559      2.839      1.606      0.108
    PRE#4            3.519      2.762      1.274      0.203
    PRE#5            4.174      2.874      1.452      0.146
    PRE#6            1.763      3.794      0.465      0.642

 POST#4   ON
    PRE#1           27.257      1.686     16.166      0.000
    PRE#2           27.323      1.553     17.595      0.000
    PRE#3           27.886      1.531     18.211      0.000
    PRE#4           27.765      1.533     18.116      0.000
    PRE#5           27.064      1.622     16.688      0.000
    PRE#6           25.225      0.000    999.000    999.000

 POST#5   ON
    PRE#1            2.889      0.567      5.095      0.000
```

```
    PRE#2              2.359      0.558      4.227      0.000
    PRE#3              2.275      0.546      4.166      0.000
    PRE#4              1.057      0.564      1.873      0.061
    PRE#5              2.471      0.637      3.881      0.000
    PRE#6              1.809      0.768      2.354      0.019

 POST#6   ON
    PRE#1              3.805      1.472      2.584      0.010
    PRE#2              3.928      1.420      2.766      0.006
    PRE#3              3.460      1.506      2.297      0.022
    PRE#4              2.764      1.421      1.945      0.052
    PRE#5              3.135      1.632      1.922      0.055
    PRE#6              4.396      1.468      2.994      0.003

Latent Class Pattern 2 1 1

 POST#1   ON
    PRE#1              3.619      0.595      6.088      0.000
    PRE#2              1.384      0.864      1.602      0.109
    PRE#3            -24.503      0.000    999.000    999.000
    PRE#4              0.966      0.699      1.382      0.167
    PRE#5              2.403      0.755      3.183      0.001
    PRE#6            -23.440      0.000    999.000    999.000

 POST#2   ON
    PRE#1             27.041      0.758     35.668      0.000
    PRE#2             27.800      0.595     46.721      0.000
    PRE#3             26.689      0.630     42.371      0.000
    PRE#4             26.403      0.625     42.213      0.000
    PRE#5             26.786      0.812     32.999      0.000
    PRE#6             27.091      0.000    999.000    999.000

 POST#3   ON
    PRE#1             33.542      1.065     31.503      0.000
    PRE#2             34.143      0.815     41.910      0.000
    PRE#3             33.812      0.799     42.305      0.000
    PRE#4             33.061      0.819     40.376      0.000
    PRE#5             33.114      1.156     28.651      0.000
    PRE#6             33.181      0.000    999.000    999.000

 POST#4   ON
    PRE#1              3.417      1.408      2.428      0.015
    PRE#2              3.498      1.351      2.590      0.010
    PRE#3              3.638      1.337      2.722      0.006
    PRE#4              4.500      1.311      3.433      0.001
    PRE#5              4.196      1.389      3.020      0.003
    PRE#6              2.282      2.187      1.043      0.297

 POST#5   ON
    PRE#1              2.334      0.470      4.970      0.000
    PRE#2              1.790      0.435      4.115      0.000
    PRE#3              1.657      0.434      3.817      0.000
```

The parameters with values less than -15 were specified in the information matrix error at the beginning of the output.

PRE#4	1.246	0.426	2.923	0.003
PRE#5	2.618	0.512	5.118	0.000
PRE#6	1.651	0.689	2.397	0.017
POST#6 ON				
PRE#1	1.825	0.733	2.491	0.013
PRE#2	1.823	0.671	2.719	0.007
PRE#3	1.211	0.711	1.702	0.089
PRE#4	1.792	0.623	2.878	0.004
PRE#5	2.153	0.778	2.766	0.006
PRE#6	-22.842	0.000	999.000	999.000

Latent Class Pattern 3 1 1

POST#1 ON				
PRE#1	2.634	0.579	4.547	0.000
PRE#2	1.643	0.634	2.591	0.010
PRE#3	1.436	0.662	2.171	0.030
PRE#4	0.824	0.654	1.260	0.208
PRE#5	0.997	0.801	1.244	0.214
PRE#6	0.795	1.113	0.715	0.475
POST#2 ON				
PRE#1	1.599	1.149	1.392	0.164
PRE#2	2.881	0.739	3.901	0.000
PRE#3	2.242	0.797	2.814	0.005
PRE#4	0.424	1.026	0.413	0.679
PRE#5	1.410	0.935	1.509	0.131
PRE#6	1.814	1.022	1.776	0.076
POST#3 ON				
PRE#1	2.219	0.966	2.297	0.022
PRE#2	1.121	1.249	0.897	0.370
PRE#3	1.586	1.003	1.581	0.114
PRE#4	1.901	0.838	2.269	0.023
PRE#5	0.031	3.843	0.008	0.994
PRE#6	-23.631	0.000	999.000	999.000
POST#4 ON				
PRE#1	0.408	1.756	0.232	0.816
PRE#2	1.048	0.947	1.107	0.268
PRE#3	1.952	0.684	2.855	0.004
PRE#4	2.246	0.625	3.594	0.000
PRE#5	2.088	0.711	2.938	0.003
PRE#6	-23.804	0.000	999.000	999.000
POST#5 ON				
PRE#1	0.809	0.903	0.895	0.371
PRE#2	1.880	0.530	3.545	0.000
PRE#3	1.776	0.547	3.249	0.001
PRE#4	1.506	0.504	2.989	0.003
PRE#5	2.470	0.491	5.030	0.000

```
    PRE#6              1.227      0.736      1.668      0.095

 POST#6   ON
    PRE#1              2.244      1.025      2.188      0.029
    PRE#2              2.216      0.864      2.566      0.010
    PRE#3              2.264      0.854      2.651      0.008
    PRE#4              1.535      0.843      1.821      0.069
    PRE#5              1.481      0.997      1.486      0.137
    PRE#6              1.504      1.558      0.965      0.334

QUALITY OF NUMERICAL RESULTS

     Condition Number for the Information Matrix           0.402E-06
       (ratio of smallest to largest eigenvalue)
```

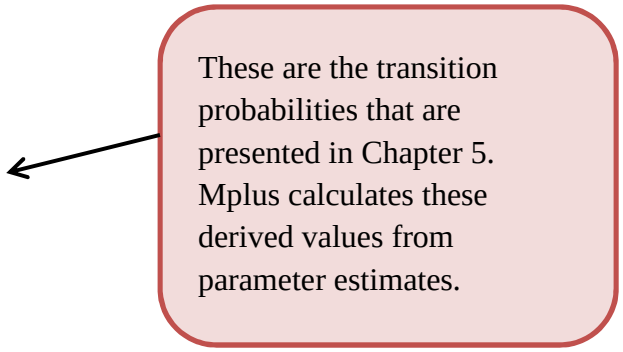
The TECH15 output converts the transition parameters into conditional probabilities.

A condition number of this magnitude does not indicate problems with identification.

```
TECHNICAL 15 OUTPUT

     ESTIMATED CONDITIONAL PROBABILITIES FOR THE CLASS VARIABLES

     P(G=1)=0.213
     P(G=2)=0.341
     P(G=3)=0.446


     P(PRE=1|G=1)=0.245
     P(PRE=2|G=1)=0.180
     P(PRE=3|G=1)=0.170
     P(PRE=4|G=1)=0.183
     P(PRE=5|G=1)=0.056
     P(PRE=6|G=1)=0.049
     P(PRE=7|G=1)=0.117


     P(PRE=1|G=2)=0.188
     P(PRE=2|G=2)=0.165
     P(PRE=3|G=2)=0.173
     P(PRE=4|G=2)=0.226
     P(PRE=5|G=2)=0.069
     P(PRE=6|G=2)=0.031
     P(PRE=7|G=2)=0.147


     P(PRE=1|G=3)=0.149
     P(PRE=2|G=3)=0.134
```

These are derived values that give the probability of pretest group membership, given treatment group membership.

```
P(PRE=3|G=3)=0.177
P(PRE=4|G=3)=0.183
P(PRE=5|G=3)=0.117
P(PRE=6|G=3)=0.054
P(PRE=7|G=3)=0.186


P(POST=1|G=1,PRE=1)=0.276
P(POST=2|G=1,PRE=1)=0.180
P(POST=3|G=1,PRE=1)=0.075
P(POST=4|G=1,PRE=1)=0.059
P(POST=5|G=1,PRE=1)=0.205
P(POST=6|G=1,PRE=1)=0.059
P(POST=7|G=1,PRE=1)=0.146


P(POST=1|G=1,PRE=2)=0.037
P(POST=2|G=1,PRE=2)=0.311
P(POST=3|G=1,PRE=2)=0.076
P(POST=4|G=1,PRE=2)=0.092
P(POST=5|G=1,PRE=2)=0.175
P(POST=6|G=1,PRE=2)=0.096
P(POST=7|G=1,PRE=2)=0.212


P(POST=1|G=1,PRE=3)=0.000
P(POST=2|G=1,PRE=3)=0.140
P(POST=3|G=1,PRE=3)=0.154
P(POST=4|G=1,PRE=3)=0.192
P(POST=5|G=1,PRE=3)=0.192
P(POST=6|G=1,PRE=3)=0.072
P(POST=7|G=1,PRE=3)=0.252


P(POST=1|G=1,PRE=4)=0.088
P(POST=2|G=1,PRE=4)=0.028
P(POST=3|G=1,PRE=4)=0.085
P(POST=4|G=1,PRE=4)=0.264
P(POST=5|G=1,PRE=4)=0.088
P(POST=6|G=1,PRE=4)=0.056
P(POST=7|G=1,PRE=4)=0.392


P(POST=1|G=1,PRE=5)=0.000
P(POST=2|G=1,PRE=5)=0.005
P(POST=3|G=1,PRE=5)=0.143
P(POST=4|G=1,PRE=5)=0.116
P(POST=5|G=1,PRE=5)=0.320
P(POST=6|G=1,PRE=5)=0.071
P(POST=7|G=1,PRE=5)=0.345
```



```
P(POST=1|G=1,PRE=6)=0.064
P(POST=2|G=1,PRE=6)=0.116
P(POST=3|G=1,PRE=6)=0.013
P(POST=4|G=1,PRE=6)=0.019
P(POST=5|G=1,PRE=6)=0.171
P(POST=6|G=1,PRE=6)=0.260
P(POST=7|G=1,PRE=6)=0.357


P(POST=1|G=1,PRE=7)=0.024
P(POST=2|G=1,PRE=7)=0.000
P(POST=3|G=1,PRE=7)=0.006
P(POST=4|G=1,PRE=7)=0.000
P(POST=5|G=1,PRE=7)=0.070
P(POST=6|G=1,PRE=7)=0.008
P(POST=7|G=1,PRE=7)=0.893


P(POST=1|G=2,PRE=1)=0.332
P(POST=2|G=2,PRE=1)=0.072
P(POST=3|G=2,PRE=1)=0.059
P(POST=4|G=2,PRE=1)=0.038
P(POST=5|G=2,PRE=1)=0.167
P(POST=6|G=2,PRE=1)=0.046
P(POST=7|G=2,PRE=1)=0.285


P(POST=1|G=2,PRE=2)=0.046
P(POST=2|G=2,PRE=2)=0.200
P(POST=3|G=2,PRE=2)=0.141
P(POST=4|G=2,PRE=2)=0.054
P(POST=5|G=2,PRE=2)=0.127
P(POST=6|G=2,PRE=2)=0.060
P(POST=7|G=2,PRE=2)=0.372


P(POST=1|G=2,PRE=3)=0.000
P(POST=2|G=2,PRE=3)=0.089
P(POST=3|G=2,PRE=3)=0.136
P(POST=4|G=2,PRE=3)=0.083
P(POST=5|G=2,PRE=3)=0.149
P(POST=6|G=2,PRE=3)=0.044
P(POST=7|G=2,PRE=3)=0.499


P(POST=1|G=2,PRE=4)=0.039
P(POST=2|G=2,PRE=4)=0.064
P(POST=3|G=2,PRE=4)=0.062
P(POST=4|G=2,PRE=4)=0.188
P(POST=5|G=2,PRE=4)=0.095
P(POST=6|G=2,PRE=4)=0.075
P(POST=7|G=2,PRE=4)=0.478
```

```
P(POST=1|G=2,PRE=5)=0.116
P(POST=2|G=2,PRE=5)=0.066
P(POST=3|G=2,PRE=5)=0.046
P(POST=4|G=2,PRE=5)=0.097
P(POST=5|G=2,PRE=5)=0.263
P(POST=6|G=2,PRE=5)=0.076
P(POST=7|G=2,PRE=5)=0.336


P(POST=1|G=2,PRE=6)=0.000
P(POST=2|G=2,PRE=6)=0.151
P(POST=3|G=2,PRE=6)=0.083
P(POST=4|G=2,PRE=6)=0.024
P(POST=5|G=2,PRE=6)=0.170
P(POST=6|G=2,PRE=6)=0.000
P(POST=7|G=2,PRE=6)=0.571


P(POST=1|G=2,PRE=7)=0.028
P(POST=2|G=2,PRE=7)=0.000
P(POST=3|G=2,PRE=7)=0.000
P(POST=4|G=2,PRE=7)=0.004
P(POST=5|G=2,PRE=7)=0.051
P(POST=6|G=2,PRE=7)=0.023
P(POST=7|G=2,PRE=7)=0.894


P(POST=1|G=3,PRE=1)=0.324
P(POST=2|G=3,PRE=1)=0.038
P(POST=3|G=3,PRE=1)=0.088
P(POST=4|G=3,PRE=1)=0.016
P(POST=5|G=3,PRE=1)=0.046
P(POST=6|G=3,PRE=1)=0.071
P(POST=7|G=3,PRE=1)=0.417


P(POST=1|G=3,PRE=2)=0.129
P(POST=2|G=3,PRE=2)=0.145
P(POST=3|G=3,PRE=2)=0.031
P(POST=4|G=3,PRE=2)=0.033
P(POST=5|G=3,PRE=2)=0.142
P(POST=6|G=3,PRE=2)=0.074
P(POST=7|G=3,PRE=2)=0.445


P(POST=1|G=3,PRE=3)=0.109
P(POST=2|G=3,PRE=3)=0.080
P(POST=3|G=3,PRE=3)=0.052
P(POST=4|G=3,PRE=3)=0.084
P(POST=5|G=3,PRE=3)=0.133
```

```
P(POST=6|G=3,PRE=3)=0.080
P(POST=7|G=3,PRE=3)=0.462


P(POST=1|G=3,PRE=4)=0.069
P(POST=2|G=3,PRE=4)=0.015
P(POST=3|G=3,PRE=4)=0.083
P(POST=4|G=3,PRE=4)=0.132
P(POST=5|G=3,PRE=4)=0.118
P(POST=6|G=3,PRE=4)=0.045
P(POST=7|G=3,PRE=4)=0.539


P(POST=1|G=3,PRE=5)=0.072
P(POST=2|G=3,PRE=5)=0.035
P(POST=3|G=3,PRE=5)=0.011
P(POST=4|G=3,PRE=5)=0.099
P(POST=5|G=3,PRE=5)=0.273
P(POST=6|G=3,PRE=5)=0.038
P(POST=7|G=3,PRE=5)=0.473


P(POST=1|G=3,PRE=6)=0.083
P(POST=2|G=3,PRE=6)=0.076
P(POST=3|G=3,PRE=6)=0.000
P(POST=4|G=3,PRE=6)=0.000
P(POST=5|G=3,PRE=6)=0.112
P(POST=6|G=3,PRE=6)=0.055
P(POST=7|G=3,PRE=6)=0.674


P(POST=1|G=3,PRE=7)=0.047
P(POST=2|G=3,PRE=7)=0.015
P(POST=3|G=3,PRE=7)=0.019
P(POST=4|G=3,PRE=7)=0.022
P(POST=5|G=3,PRE=7)=0.041
P(POST=6|G=3,PRE=7)=0.015
P(POST=7|G=3,PRE=7)=0.840



Beginning Time:  20:03:17
   Ending Time:  23:22:00
  Elapsed Time:  03:18:43
```

The estimation of this model took 3 hours for a 24 parallel processor computer.

MUTHEN & MUTHEN
3463 Stoner Ave.
Los Angeles, CA 90066

Tel: (310) 391-9971

```
Fax: (310) 391-8971
Web: www.StatModel.com
Support: Support@StatModel.com

Copyright (c) 1998-2014 Muthen & Muthen
```

References

Abar, B., & Loken, E. (2012). Consequences of Fitting Nonidentified Latent Class Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 1–15. http://doi.org/10.1080/10705511.2012.634701

Agresti, A. (2012). *Categorical Data Analysis* (3 edition). Hoboken, NJ: Wiley.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389–421. http://doi.org/10.1002/sce.20303

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, *39*(10), 952–978. http://doi.org/10.1002/tea.10053

Asparouhov, T. & Muthén, B. (2012). Mplus WebNote 14: Using Mplus TECH11 and TECH14 to test the number of latent classes. Available from www.statmodel.com.

Bailey, J. M. (2007). Development of a Concept Inventory to Assess Students' Understanding and Reasoning Difficulties about the Properties and Formation of Stars. *Astronomy Education Review*, *6*(2), 133–139. http://doi.org/10.3847/AER2007028

Balint, T. A., Teodorescu, R., Colvin, K., Choi, Y.-J., & Pritchard, D. E. (2015). Comparing Measures of Student Performance in Hybrid and MOOC Physics Courses. *European Journal of Physics Education*, *6*(3). http://doi.org/10.20308/ejpe.77043

Bao, L., & Redish, E. F. (2006). Model analysis: Representing and assessing the dynamics of student learning. *Physical Review Special Topics - Physics Education Research*, *2*(1). http://doi.org/10.1103/PhysRevSTPER.2.010103

Berzofsky, M. E., Biemer, P. P., & Kalsbeek, W. D. (2014). Local Dependence in Latent Class Analysis of Rare and Sensitive Events. *Sociological Methods & Research, 43*(1), 137–170. http://doi.org/10.1177/0049124113506407

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. http://doi.org/10.1080/0969595980050102

Bradshaw, L., & Templin, J. (2014). Combining Item Response Theory and Diagnostic Classification Models: A Psychometric Model for Scaling Ability and Diagnosing Misconceptions. *Psychometrika*, *79*(3), 403–425. http://doi.org/10.1007/s11336-013-9350-4

Brennan, R. L. (2010). Evidence-Centered Assessment Design and the Advanced Placement Program®: A Psychometrician's Perspective. *Applied Measurement in Education, 23*(4), 392–401. http://doi.org/10.1080/08957347.2010.510973

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic Assessment With Ordered Multiple-Choice Items. *Educational Assessment*, *11*(1), 33–63. http://doi.org/10.1207/s15326977ea1101_2

Bye, B. V., & Schechter, E. S. (1986). A Latent Markov Model Approach to the Estimation of Response Errors in Multiwave Panel Data. *Journal of the American Statistical Association, 81*(394), 375–380. http://doi.org/10.1080/01621459.1986.10478281

Cahyadi, V. (2004). The effect of interactive engagement teaching on student understanding of introductory physics at the faculty of engineering, University of Surabaya, Indonesia. *Higher Education Research & Development*, *23*(4), 455–464. http://doi.org/10.1080/0729436042000276468

Champagne, A. B., Klopfer, L. E., & Anderson, J. H. (1980). Factors influencing the learning of classical mechanics. *American Journal of Physics, 48*(12), 1074–1079. http://doi.org/10.1119/1.12290

Cheng, K. K., Thacker, B. A., Cardenas, R. L., & Crouch, C. (2004). Using an online homework system enhances students' learning of physics concepts in an introductory physics course. *American Journal of Physics, 72*(11), 1447–1453. http://doi.org/10.1119/1.1768555

Cho, S.-J., Bottge, B. A., Cohen, A. S., & Kim, S.-H. (2011). Detecting Cognitive Change in the Math Skills of Low-Achieving Adolescents. *The Journal of Special Education, 45*(2), 67–76. http://doi.org/10.1177/0022466909351579

Cho, S.-J., Cohen, A. S., & Bottge, B. (2013). Detecting Intervention Effects Using a Multilevel Latent Transition Analysis with a Mixture IRT Model. *Psychometrika, 78*(3), 576–600. http://doi.org/10.1007/s11336-012-9314-0

Cho, S.-J., Cohen, A. S., Kim, S.-H., & Bottge, B. (2010). Latent Transition Analysis With a Mixture Item Response Theory Measurement Model. *Applied Psychological Measurement, 34*(7), 483–504. http://doi.org/10.1177/0146621610362978

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics, 50*(1), 66–71. http://doi.org/10.1119/1.12989

Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics, 73*(12), 1172. http://doi.org/10.1119/1.2117109

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: with applications in the social behavioral, and health sciences*. Hoboken, N.J: Wiley.

Collins, L. M., & Wugalter, S. E. (1992). Latent Class Models for Stage-Sequential Dynamic Latent Variables. *Multivariate Behavioral Research*, *27*(1), 131–157. http://doi.org/10.1207/s15327906mbr2701_8

Crouch, C. H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics*, *69*(9), 970–977. http://doi.org/10.1119/1.1374249

Cummings, K., Marx, J., Thornton, R., & Kuhl, D. (1999). Evaluating innovation in studio physics. *American Journal of Physics*, *67*(S1), S38–S44. http://doi.org/10.1119/1.19078

Dancy, M. H., & Beichner, R. (2006). Impact of animation on assessment of conceptual understanding in physics. *Physical Review Special Topics - Physics Education Research*, *2*(1), 10104. http://doi.org/10.1103/PhysRevSTPER.2.010104

Davenport, G.A. (2008) *The reliability of the Force and Motion Conceptual Evaluation.* (Unpublished Master's thesis). University of Maine, Orono, ME.

Davenport, G.A. (2012, October). *Latent Class Analysis as a method for scoring concept inventory instruments*. Paper presented at the Northeastern Educational Research Association Annual Meeting. Rocky Hill, CT.

Davenport, G.A., Rogers, H.J. (April 2013). *Strategic Measurement Model Selection for Conceptual Diagnostic Assessments*. American Education Research Association annual conference. San Francisco, CA.

Davenport, G.D. (2014). *The Use of Concept Inventory Data in Advanced Placement Classrooms.* Unpublished report. College Board Graduate Research Fellowship Program.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-Variable Latent-Class Models. *Journal of the American Statistical Association*, *83*(401), 173–178. http://doi.org/10.2307/2288938

de la Torre, J., & Minchen, N. (2014). Cognitively Diagnostic Assessments and the Cognitive Diagnosis Model Framework. *Psicología Educativa*, *20*(2), 89–97. http://doi.org/10.1016/j.pse.2014.11.001

DeMars, C. E. (2013). A Tutorial on Interpreting Bifactor Model Scores. *International Journal of Testing*, *13*(4), 354–378. http://doi.org/10.1080/15305058.2013.799067

Demastes, S. S., Good, R. G., & Peebles, P. (1996). Patterns of conceptual change in evolution. *Journal of Research in Science Teaching*, *33*(4), 407–431. http://doi.org/10.1002/(SICI)1098-2736(199604)33:4<407::AID-TEA4>3.0.CO;2-W

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Dietz, R. D., Pearson, R. H., Semak, M. R., & Willis, C. W. (2012). Gender bias in the force concept inventory? In *AIP Conference Proceedings* (Vol. 1413, pp. 171–174). AIP Publishing. http://doi.org/10.1063/1.3680022

Lin, D., & Robert, B. (2007). *Designing An Energy Assessmen tto Evaluate Student Understanding of Energy Topics* (Doctoral dissertation, North Carolina State University).

Ding, L., Chabay, R., & Sherwood, B. (2013). How do students in an innovative principle-based mechanics course understand energy concepts? *Journal of Research in Science Teaching*, *50*(6), 722–747. http://doi.org/10.1002/tea.21097

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Special Topics - Physics Education Research*, *2*(1), 10105. http://doi.org/10.1103/PhysRevSTPER.2.010105

Disessa, A. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. Stevens (Eds.), *Mental Models* (pp. 15–34). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Disessa, A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), *Constructivism in the computer age* (pp. 49–70). Lawrence Erlbaum. Retrieved from http://www.questia.com/PM.qst?a=o&d=13634588#

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching*, *46*(6), 606–609. http://doi.org/10.1002/tea.20316

Formann, A. K. (1982). Linear Logistic Latent Class Analysis. *Biometrical Journal*, *24*(2), 171–190. http://doi.org/10.1002/bimj.4710240209

Fulmer, G. W., Liang, L. L., & Liu, X. (2014). Applying a Force and Motion Learning Progression over an Extended Time Span Using the Force Concept Inventory. *International Journal of Science Education*, *36*(17), 2918–2936.

Gierl, M. J., & Cui, Y. (2008). Defining Characteristics of Diagnostic Classification Models and the Problem of Retrofitting in Cognitive Diagnostic Assessment. *Measurement: Interdisciplinary Research and Perspectives*, *6*(4), 263–268. http://doi.org/10.1080/15366360802497762

Glasersfeld, E. von. (1998). Cognition, Construction of Knowledge, and Teaching. In M. R. Matthews (Ed.), *Constructivism in Science Education* (pp. 11–30). Springer Netherlands. Retrieved from http://link.springer.com/chapter/10.1007/978-94-011-5032-3_2

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231. http://doi.org/10.1093/biomet/61.2.215

Hagenaars, J. A. (1998). Categorical Causal Modeling Latent Class Analysis and Directed Log-Linear Models with Latent Variables. *Sociological Methods & Research*, *26*(4), 436–486. http://doi.org/10.1177/0049124198026004002

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64–74. http://doi.org/10.1119/1.18809

Halloun, I., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics*, *53*(11), 1043–1055. http://doi.org/10.1119/1.14030

Hammer, D. (2000). Student resources for learning introductory physics. *American Journal of Physics*, *68*(S1), S52–S59. http://doi.org/10.1119/1.19520

Hammer, D., & Elby, A. (2003). Tapping Epistemological Resources for Learning Physics. *The Journal of the Learning Sciences*, *12*(1), 53–90.

Haslam, F., & Treagust, D. F. (1987). Diagnosing secondary students' misconceptions of photosynthesis and respiration in plants using a two-tier multiple choice instrument. *Journal of Biological Education*, *21*(3), 203–211. http://doi.org/10.1080/00219266.1987.9654897

Heller, P., & Huffman, D. (1995). Interpreting the force concept inventory: A reply to Hestenes and Halloun. *The Physics Teacher*, *33*(8), 503–503. http://doi.org/10.1119/1.2344279

Henderson, C. (2002). Common Concerns About the Force Concept Inventory. *The Physics Teacher*, *40*(9), 542–547. http://doi.org/10.1119/1.1534822

Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller. *The Physics Teacher*, *33*(8), 502–502. http://doi.org/10.1119/1.2344278

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141–158. http://doi.org/10.1119/1.2343497

Hoellwarth, C., Moelter, M. J., & Knight, R. D. (2005). A direct comparison of conceptual learning and problem solving ability in traditional and studio style classrooms. *American Journal of Physics*, *73*(5), 459. http://doi.org/10.1119/1.1862633

Huff, K., & Goodman, D. (1997). The demand for cognitive diagnostic assessment. In Leighton, Jacqueline P. & Gierl, Mark J. (Eds.), *Cognitive Diagnostic Assessment for Education* (pp. 19–60). New York, NY: Cambridge University Press.

Huffman, D., & Heller, P. (1995). What does the force concept inventory actually measure? *The Physics Teacher*, *33*(3), 138–143. http://doi.org/10.1119/1.2344171

Huang, C., & Mislevy, R. J. (2010). An application of the polytomous Rasch model to mixed strategies. *Handbook of polytomous item response theory models*, 211-228.

Kane, M. (2011). Validating Score Interpretations and Uses. *Language Testing*, *29*(1), 3–17. http://doi.org/10.1177/0265532211417210

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. http://doi.org/10.1037/0033-2909.112.3.527

Kim, J.-S. (2005). A Latent-Change Scaling Model for Longitudinal Multiple Choice Data. *Multivariate Behavioral Research*, *40*(1), 53–82. http://doi.org/10.1207/s15327906mbr4001_3

Koehler, K. J. (1986). Goodness-of-Fit Tests for Log-Linear Models in Sparse Contingency Tables. *Journal of the American Statistical Association*, *81*(394), 483–493. http://doi.org/10.1080/01621459.1986.10478294

Kost, L. E., Pollock, S. J., & Finkelstein, N. D. (2009). Characterizing the gender gap in introductory physics. *Physical Review Special Topics - Physics Education Research*, *5*(1), 10101. http://doi.org/10.1103/PhysRevSTPER.5.010101

Lasry, N., Rosenfield, S., Dedic, H., Dahan, A., & Reshef, O. (2011). The puzzling reliability of the Force Concept Inventory. *American Journal of Physics*, *79*(9), 909–912. http://doi.org/10.1119/1.3602073

Lazarsfeld, P.F., & Henry, N.W. (1968). Latent structure analysis. Houghton Mifflin, Boston.

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

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767–778. http://doi.org/10.1093/biomet/88.3.767

Louca, L., Elby, A., Hammer, D., & Kagey, T. (2004). Epistemological Resources: Applying a New Epistemological Framework to Science Instruction. *Educational Psychologist*, *39*(1), 57–68. http://doi.org/10.1207/s15326985ep3901_6

Maloney, D. P., O'Kuma, T. L., Hieggelke, C. J., & Heuvelen, A. V. (2001). Surveying students' conceptual knowledge of electricity and magnetism. *American Journal of Physics*, *69*(S1), S12–S23. http://doi.org/10.1119/1.1371296

Marshall, J. A., Hagedorn, E. A., & O'Connor, J. (2009). Anatomy of a physics test: Validation of the physics items on the Texas Assessment of Knowledge and Skills. *Physical Review Special Topics - Physics Education Research*, *5*(1), 10104. http://doi.org/10.1103/PhysRevSTPER.5.010104

McCloskey, M., & Kohl, D. (1983). Naive physics: The curvilinear impetus principle and its role in interactions with moving objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(1), 146–156. http://doi.org/10.1037/0278-7393.9.1.146

McCullough, L. (2011, July 19-20). Gender Differences in Student Responses to Physics Conceptual Questions Based on Question Context. Paper presented at ASQ Advancing the STEM Agenda in Education, the Workplace and Society, University of Wisconsin-Stout. Retrieved May 13, 2016, from http://rube.asq.org/edu/2011/06/continuous-improvement/gender-differences-in-student-responses-to-physics-conceptual-questions-based-on-question-content.pdf

McDermott, L. C. (1991). Millikan Lecture 1990: What we teach and what is learned—Closing the gap. *American Journal of Physics*, *59*(4), 301–315. http://doi.org/10.1119/1.16539

McDermott, L. C., & Redish, E. F. (1999). Resource Letter: PER-1: Physics Education Research. *American Journal of Physics*, *67*(9), 755–767. http://doi.org/10.1119/1.19122

McDermott, L. C., Rosenquist, M. L., & Zee, E. H. van. (1987). Student difficulties in connecting graphs and physics: Examples from kinematics. *American Journal of Physics*, *55*(6), 503–513. http://doi.org/10.1119/1.15104

McDermott, L. C., & Shaffer, P. S. (2001). *Tutorials in Introductory Physics*. Upper Saddle River, N.J.: Prentice Hall College Div.

McLachlan, G., & Peel, D. (2004). *Finite Mixture Models*. John Wiley & Sons.

Mislevy, R. J., & Haertel, G. D. (2007). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20. http://doi.org/10.1111/j.1745-3992.2006.00075.x

Muthen, B.O. (Sept 9, 2002). Reporting Results. Message posted to http://www.statmodel.com/discussion/messages/13/173.html

Muthen, B.O. (May 5, 2006). First order derivative product matrix error. Message posted to http://www.statmodel.com/cgi-bin/discus/discus.cgi?pg=prev&topic=13&page=336

Muthen, B.O. (Feb 3, 2007a). Latent Transition Analysis (LTA). Message posted to http://www.statmodel2.com/discussion/messages/13/278.html?1206110600

Muthen, B.O. (Aug 31, 2007b). First order derivative product matrix error. Message posted to http://www.statmodel.com/cgi-bin/discus/discus.cgi?pg=prev&topic=13&page=336

Muthen, B.O. (Nov 12, 2008). Ordinal, Nominal, and Interval Indicators. Message posted to http://www.statmodel.com/discussion/messages/13/3720.html?1432342065

Muthén, B.O. & Asparouhov, T. (2011). Mplus Web Notes 13: LTA in Mplus: Transition probabilities influenced by covariates. Available from www.statmodel.com.

Muthen, B.O., & Muthen, L.K. (2009a). Topic 5 Categorical latent variable modeling with cross-sectional data [Lecture video, recorded at Freie University, Berlin]. Retrieved from www.statmodel.com/course_materials.shtml.

Muthen, B.O., & Muthen, L.K. (2009b). Topic 6 Categorical latent variable modeling with longitudinal data [Lecture video, recorded at Freie University, Berlin]. Retrieved from www.statmodel.com/course_materials.shtml.

Muthén, L.K. and Muthén, B.O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. http://doi.org/10.1080/10705510701575396

O'connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32*(3), 396–402. http://doi.org/10.3758/BF03200807

Pollock, S. J., Finkelstein, N. D., & Kost, L. E. (2007). Reducing the gender gap in the physics classroom: How sufficient is interactive engagement? *Physical Review Special Topics - Physics Education Research*, *3*(1), 10107. http://doi.org/10.1103/PhysRevSTPER.3.010107

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227. http://doi.org/10.1002/sce.3730660207

Ramlo, S. (2008). Validity and reliability of the force and motion conceptual evaluation. *American Journal of Physics*, *76*(9), 882–886. http://doi.org/10.1119/1.2952440

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York, NY: Springer New York. Retrieved from http://dx.doi.org/10.1007/978-1-4612-4578-0

Redish, E. F. (1999). Millikan Lecture 1998: Building a Science of Teaching Physics. *American Journal of Physics*, *67*(7), 562–573. http://doi.org/10.1119/1.19326

Redish, E. F., Saul, J. M., & Steinberg, R. N. (1997). On the effectiveness of active-engagement microcomputer-based laboratories. *American Journal of Physics*, *65*(1), 45–54. http://doi.org/10.1119/1.18498

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

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*,

35(3), 265–296. http://doi.org/10.1002/(SICI)1098-2736(199803)35:3<265::AID-TEA3>3.0.CO;2-P

Sadler, P. M., Coyle, H., Miller, J. L., Cook-Smith, N., Dussault, M., & Gould, R. R. (2009). The Astronomy and Space Science Concept Inventory: Development and Validation of Assessment Instruments Aligned with the K–12 National Science Standards. *Astronomy Education Review*, *8*(1). http://doi.org/10.3847/AER2009024

Sahin, M. (2009). Effects of Problem-Based Learning on University Students' Epistemological Beliefs About Physics and Physics Learning and Conceptual Understanding of Newtonian Mechanics. *Journal of Science Education and Technology*, *19*(3), 266–275. http://doi.org/10.1007/s10956-009-9198-7

Savinainen, A., & Scott, P. (2002). The Force Concept Inventory: a tool for monitoring student learning. *Physics Education*, *37*(1), 45. http://doi.org/10.1088/0031-9120/37/1/306

Schneider, M., & Hardy, I. (2013). Profiles of Inconsistent Knowledge in Children's Pathways of Conceptual Change. *Developmental Psychology*, *49*(9), 1639–1649.

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The Genetics Concept Assessment: a new concept inventory for gauging student understanding of genetics. *CBE Life Sciences Education*, *7*(4), 422–430. http://doi.org/10.1187/cbe.08-08-0045

Smith, T. I., & Wittmann, M. C. (2007). Comparing three methods for teaching Newton's third law. *Physical Review Special Topics - Physics Education Research*, *3*(2), 20105. http://doi.org/10.1103/PhysRevSTPER.3.020105

Smith, T. I., & Wittmann, M. C. (2008). Applying a resources framework to analysis of the Force and Motion Conceptual Evaluation. *Physical Review Special Topics - Physics Education Research*, *4*(2), 20101. http://doi.org/10.1103/PhysRevSTPER.4.020101

Sokoloff, D. R., & Thornton, R. K. (2004). *Interactive Lecture Demonstrations, Active Learning in Introductory Physics* (1 edition). New York: Wiley.

Sokoloff, D. R., Thornton, R. K., & Laws, P. W. (2011). *RealTime Physics Active Learning Laboratories, Module 1: Mechanics* (3 edition). Hoboken, N.J.: Wiley.

Springuel, R.P. (2010). Applying Cluster Analysis to Physics Education Research Data (Unpublished doctoral dissertation). University of Maine, Orono, ME.

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching*, *46*(6), 699–715. http://doi.org/10.1002/tea.20308

Stone, A., Allen, K., Rhoads, T. R., Murphy, T. J., Shehab, R. L., & Saha, C. (2003). The statistics concept inventory: a pilot study. In *Frontiers in Education, 2003. FIE 2003 33rd Annual* (Vol. 1, p. T3D–1–6 Vol.1). http://doi.org/10.1109/FIE.2003.1263336

Stone, C. A., Feifei Ye, Xiaowen Zhu, & Lane, S. (2010). Providing Subscale Scores for Diagnostic Information: A Case Study When the Test is Essentially Unidimensional. *Applied Measurement in Education*, *23*(1), 63–86. http://doi.org/10.1080/08957340903423651

Thornton, R. K., Kuhl, D., Cummings, K., & Marx, J. (2009). Comparing the force and motion conceptual evaluation and the force concept inventory. *Physical Review Special Topics - Physics Education Research*, *5*(1), 10105. http://doi.org/10.1103/PhysRevSTPER.5.010105

Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula. *American Journal of Physics*, *66*(4), 338–352. http://doi.org/10.1119/1.18863

Toulmin, S. E. (2003). *The Uses of Argument* (Updated edition). Cambridge, U.K. ; New York: Cambridge University Press.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction*, *4*(1), 45–69. http://doi.org/10.1016/0959-4752(94)90018-3

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*(4), 535–585. http://doi.org/10.1016/0010-0285(92)90018-W

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, *57*(2), 307–333. http://doi.org/10.2307/1912557

Wallace, C. S., & Bailey, J. M. (2010). Do Concept Inventories Actually Measure Anything? *Astronomy Education Review*, *9*(1). http://doi.org/10.3847/AER2010024

Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics*, *78*(10), 1064–1070. http://doi.org/10.1119/1.3443565

Wilson, M. (2008). Cognitive diagnosis using item response models. *Zeitschrift Für Psychologie/Journal of Psychology*, *216*(2), 74–88. http://doi.org/10.1027/0044-3409.216.2.74

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716–730. http://doi.org/10.1002/tea.20318

Young, M. A. (1982). Evaluating diagnostic criteria: A latent class paradigm. *Journal of Psychiatric Research*, *17*(3), 285–296. http://doi.org/10.1016/0022-3956(82)90007-3