ABSTRACT

Title of Dissertation: PSYCHOMETRIC ANALYSES BASED ON EVIDENCE-CENTERED DESIGN AND COGNITIVE SCIENCE OF LEARNING TO EXPLORE STUDENTS' PROBLEM-SOLVING IN PHYSICS

Chun-Wei Huang, Doctor of Philosophy, 2003

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

Most analyses of physics assessment tests have been done within the framework of classical test theory in which only the number of correct answers is considered in the scoring. More sophisticated analyses have been developed recently by physics researchers to further study students' conceptions/misconceptions in physics learning to improve physics instruction. However, they are not connected with the well-developed psychometric machinery.

The goal of this dissertation is to use a formal psychometric model to study students' conceptual understanding in physics (in particular, Newtonian mechanics). The perspective is based on the evidence-centered design (ECD) framework, building on

previous analyses of the cognitive processes of physics problem-solving and the task design from two physics tests (Force Concept Inventory, FCI and Force Motion Concept Evaluation, FMCE) that are commonly used to measure students' conceptual understanding about force-motion relationships.

Within the ECD framework, the little-known Andersen/Rasch (AR) multivariate IRT model that can deal with mixtures of strategies within individuals is then introduced and discussed, including the issue of identification of the model. To demonstrate its usefulness, four data sets (one from FCI and three from FMCE) were used and analyzed with the AR model using a Markov Chain Monte Carlo estimation procedure, carried out with the *BUGS* computer program.

Results from the first three data sets (questions were used to assess students' understanding about force-motion relationships) indicate that most students are in a mixed model state (i.e., in a transition toward understanding Newtonian mechanics) after one semester of physics learning. In particular, they incorrectly tend to believe that there must be a force acting on an object to maintain its movement, one of the common misconceptions indicated in physics literature. Findings from the last data set (which deals with acceleration) indicate that although students have improved their understanding about acceleration after one semester of instruction, they may still find it difficult to represent their understanding in terms of acceleration-time graphs. This is especially so when the object is slowing down or moving toward the left, in which case the sign of acceleration in both task scenarios is negative.

PSYCHOMETRIC ANALYSES BASED ON EVIDENCE-CENTERED DESIGN AND COGNITIVE SCIENCE OF LEARNING TO EXPLORE STUDENTS' PROBLEM-SOLVING IN PHYSICS

by

Chun-Wei Huang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2003

Advisory Committee:
Professor Robert J. Mislevy, Chair
Professor Roger Azevedo
Professor C. Mitchell Dayton
Professor Edward F. Redish
Professor James Roberts

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF FIGURES

Chapter I

Introduction

An emerging development in educational assessment is the integration of the science of learning and the science of measurement (National Research Council, 2001). Rather than separate and disconnected phases of item writing, analysis, and interpretation, researchers are seeking to coordinate psychology, task design, and psychometric analysis (Embretson, 1985, 1998). This dissertation is meant to provide an example of this line of research. It concerns innovative psychometric modeling approaches to a particular kind of assessment designed to reveal students' conceptions and misconceptions about Newtonian mechanics in physics.

It was not until 1980's that physics educators started to probe students' conceptual understanding in physics. In her review about research on conceptual understanding in mechanics (e.g., gravitational force, velocity and acceleration, and force and motion), McDermott (1984) pointed out some interesting and unexpected results from several studies. Studies about "passive" forces (e.g., the tension in a string) indicated that students, regardless of ages, have the same conceptual difficulty understanding those forces, yet most physics instructors proceed as if the concept of a passive force is easily understood (e.g., Minstrell, 1982; Sjøberg & Lie, 1981). A study of gravity showed that the serious misconceptions may sometimes lead to correct answers (Champagne, Klopfer, & Anderson, 1980). A study in velocity and acceleration revealed that students with greater facility with mathematics do not necessarily have a deeper conceptual understanding than those who have less training in mathematics (Whitaker, 1983). Studies

in the relation between force and motion showed that many students (even they had previously taken physics) have a well-integrated system of beliefs about the behavior of objects in motion – an object will not stop moving unless the initial force acting on the subject is "used up", the "Aristotelian" or "medieval" belief – that is in conflict with the Newtonian view (McCloskey, Caramazza, & Green, 1980). To further study this phenomenon, Viennot (1979) constructed a model for how students think about force, assuming that they often hold simultaneously both Newtonian and non-Newtonian conceptions of force, and the particulars of the task presented to them determine which belief system will be used for problem-solving. Finally, Lawson's study (1984) of the ability of college students to apply the work-energy and impulse-momentum relations to an actual physical system showed that a majority of students (including honors physics students and non-calculus students) could not make the necessary connection between the algebraic expressions and a real-life demonstration even though they had little difficulty in applying these relations to standard textbook problems.

Since these findings were based on different physics topics and experiments (or tasks), McDermott further suggested the following characteristics that should be considered when conducting research on conceptual understanding about physics and interpreting the results: (1) nature of instrument used to assess understanding, (2) degree of interaction between student and investigator, (3) depth of probing, (4) form of data, (5) physical setting, (6) time frame, and (7) goals of investigator.

In fact, these considerations are also important in assessing students' learning in other fields of study. In order to carry out this type of assessment, covering all of above concerns, a systematic approach that describes characteristics and implementation of each

stage in assessment would be preferred. In this dissertation, the "evidence-centered" assessment design (ECD) developed by Mislevy, Steinberg, and Almond (2003) is used. Within the ECD approach, the focus in this presentation is on the stage of the "Conceptual Assessment Framework (CAF)". The CAF concerns the interplay among three models, namely, the student model, the task model, and the evidence model. These three models correspond to three key elements (cognition, observation, and interpretation – the assessment triangle) underlying any assessment, as discussed in the recent publication entitled *Knowing What Students Know* (National Research Council, 2001, p 44). More thorough reviews on ECD as well as recent developments in cognitive science of learning (particularly in physics – e.g., Bao & Redish, 2004; diSessa, 1982; Reiner, Slotta, Chi, & Resnick, 2000) are given in the next chapter.

Investigations from research reviewed by McDermott, on one hand, provide resources for improving physics instruction to help students to learn abstract physics concepts, especially those found to be more difficult for many students. On the other hand, they lead to the issue about the statistical analysis of students' conceptual understanding in physics. Data from the above-mentioned studies was mainly collected and analyzed by laboratory observation and/or interviews, and descriptive analysis was conducted for some of the research studies in which a survey or a written test was administered to students. The latter approach (i.e., using written tasks) became more common as some useful instruments were developed – for example, the Force Concept Inventory (FCI), a well-designed research-based multiple-choice instrument developed to measure students' conceptual understanding of Newtonian mechanics (Hestenes, Wells, & Swackhammer, 1992) and the Force Motion Concept Evaluation (FMCE), designed to measure students' understanding

about force-motion relationships (Thornton & Sokoloff, 1998). Hake (2002) pointed out that most of the analyses of physics assessment tests have been done within the framework of "classical test theory" in which only the number of correct answers is considered in the scoring. This includes the commonly-used index called "normalized gain" (g), based on the difference between pre- and post-test and defined as Gain/[Gain(maximum possible)]. For example, if a class averaged 50% on the pretest, and 70% on the posttest then the class-average normalized gain = (70%-50%)/(100%-50%) = .4. Test reliability indices such as Cronbach's alpha and the Kuder-Richardson reliability coefficient (also known as KR-20) have been employed as well. These analyses mainly tell us students' mastery levels (i.e., how much they understand or how much knowledge they have gained through instruction) but shed little light on students' thinking in terms of how they respond to the tasks presented to them.

More sophisticated analyses have been developed to further study students' conceptions/misconceptions in physics learning (e.g., Bao & Redish, 2001 & 2004). Rather than examining students' responses in terms of correctness, Bao and Redish focus on how students' responses on multiple-choice questions are distributed, to explore if students have common correct or incorrect models for problem-solving. They call this *Concentration Analysis*. In addition, the second method they proposed, called *Model Analysis*, can be used to extract the probability states of students' use of different models. *Model Analysis* (summarized here in Chapter II) adapts techniques from quantum mechanics to characterize students in terms of their propensities of responding according to different perspectives on a class of physical phenomena. An analysis based on these two

methods provides much more information than the traditional analysis mentioned above in terms of exploring students' conceptual understanding about physics concepts.

However, there are some limitations to these methods (mainly, they are not well connected with current development in psychometrics), and their usefulness has not been examined in other learning subjects. This may not be an issue for physics educators, but it would be a concern for researchers in educational measurement.

For the past twenty years or so, psychometric analyses based on modern test theory (e.g., item response theory, IRT) have been well-developed, and some of them may be useful for cognitively-relevant assessment (see Junker, 1999, and National Research Council, 2001, Chapter 4, "Contributions of Measurement and Statistical Modeling to Assessment"). It is the goal of this presentation to examine one way in which a psychometric analysis can offer some advantages in studying students' conceptual understanding in physics.

In the current study, the little-known Andersen/Rasch (AR) multivariate IRT model (Andersen, 1973 & 1995) that can deal with mixtures of strategies within individuals and is parallel to Bao and Redish's *Model Analysis* is introduced, and is applied to several data sets collected by Bao and Redish. The use of the AR model is appropriate given the considerations of the CAF and the assessment triangle: It coheres nicely with the psychological perspective for the knowledge being assessed (in this study, students are viewed as in a mixed-model state for solving a physics task) and the observations (the data were collected from students' responses on items designed to reflect the mixtures of strategies within students).

In addition, other IRT models (in particular, the three-parameter logistic model, 3PL, which is commonly used in the education measurement) and latent class (LC) models are included in the review to discuss the interplay among substantive perspectives, task design, measurement models, and inferences about student learning in the ECD framework. However, they are not included in the data analysis to contrast with the AR model. This is because that the comparisons of model fit among those models are both complicated and off the main track of our objective. For example, one cannot compare 3PL with the AR model when the AR is being used to model more than two response types since the collapsing of response categories are different; that is, dichotomous vs. polytomous. Therefore, the comparisons among those models offered here are not empirical. Rather, in the next chapter we discuss their similarities and dissimilarities in terms of the way they model students' responses with respect to the psychological perspective each model represents and the purpose of assessment.

Finally, a Markov chain Monte Carlo (MCMC; Gelman, Carlin, Stern, & Rubin, 1995) estimation procedure is used here to estimate the item and person parameters under the AR model. This procedure is growing rapidly in popularity in the statistical literature due to its flexibility in model fitting. It has been applied in the context of IRT (as reviewed in Chapter II). It is of interest to explore how it performs with the AR and the data used in the current study.

In sum, this dissertation integrates ideas from the following areas of current research:

- "Evidence-centered" assessment design.
- The psychology of science learning in physics.
- Latent-trait measurement models (in particular, the AR model).
- MCMC estimation.

Each area is reviewed in detail in the following chapter. The contribution of this study is threefold. First of all, it provides an example of doing an assessment by integrating the fields of research listed above. Second, since applications of the AR model are hardly found in the psychometrics literature, this study would represent a useful contribution to educational measurement as well. Third, the findings from the current study would also be beneficial to members of the physics education community who are interested in promoting the physics instruction by providing a tool for analyzing data from tests such as FCI and FCME. This type of the analysis addresses the inferences about thinking Bao and Redish and others are interested in, and inherits the advantages of model fitting, model interpretation, and model criticism techniques of probability-based reasoning.

Chapter II

Literature Review

Assessment Design

Educational assessment concerns inference about students' ability or knowledge based on evidence elicited by given tasks (Mislevy, 1994). For example, to assess international students' English language proficiency, a TOEFL test (consisting of a collection of language tasks) can be used. Based on elements we believe are important in language proficiency for non-English speakers coming to study in the U.S., four subtests are constructed and included in the TOEFL test: listening, vocabulary and grammar, reading comprehension, and writing. Succeeding on these tasks (evidence) in TOEFL indicates the examinee possesses language proficiency (an inference we make). By using this example, we can see that an assessment design needs to address the following questions: What do you want to make inferences about (e.g., language proficiency), what do you need to see (e.g., performance with what qualities), and what features in tasks evoke the evidence you need (e.g., identifying the topic of an oral conversation in TOEFL) (Messick, 1994). In this dissertation, the target inference is the mixtures of conceptions (some correct, some erroneous) that students bring to bear on mechanics problems, with evidence coming from tasks designed to reveal those conceptions.

Evidence-centered assessment design (ECD) (Mislevy, et al., 2003) is a formal framework for designing assessments from this evidentiary-reasoning perspective. The ECD framework is introduced below to provide more details and thorough discussion in terms of the structure of an assessment design.

There are four stages in the ECD: Domain Analysis, Domain Modeling, Conceptual Assessment Framework, and Operational Assessment. Domain Analysis concerns gathering information about how people acquire knowledge or skills, and how they use them. This information is essential in assessment since it will help the assessment designer to know, for instances, under what situations we can see people doing the kinds of things and using the kinds of knowledge related to assessment. This analysis can provide clues about important features of performance situations. This information then is organized in terms of design objects called paradigms in the second stage of the ECD, Domain Modeling. There are three paradigms: the proficiency paradigms are the structures that organize potential claims about aspects of proficiency for students; the evidence paradigms state the kinds of things student might say or do that would provide evidence about these proficiencies; and the task paradigms are the kinds of situations that might evoke the evidence we need to see. At this stage, by knowing the interrelationship among these three paradigms, one starts to rough out the structure of an assessment that will be needed for future operational assessment.

In this dissertation, the required information for Domain Analysis and Domain Modeling comes from the domains of physics, physics learning, and cognitive psychology. Since the physics tasks and data used in the current study were built on Bao and Redish's expertise and familiarity with the domain (see the Chapter III for more details about the description of data sets), much of this work has been done already. Our task here is to map their thinking into the Domain Modeling structures in the course of a literature review. These are conceptual foundations from which specific assessments, and the arguments they embody, are developed. These paradigms can be viewed as narrative forms of the

evidentiary-reasoning arguments that underlie operational assessments. They ground the specifics of task authoring, scoring rules, and statistical models.

The next stage of the design, then, is the Conceptual Assessment Framework (CAF). The CAF specifies the more technical elements of an operational assessment, including, importantly, measurement models. As mentioned in the introduction, this dissertation research is mainly based on this stage, as we develop, fit, and compare three models under the AR model, and discuss how the other two psychometric models differ from the AR model in reflecting views of mechanics knowledge with respect to the Bao and Redish data sets.

The CAF consists of three major models that coordinate an assessment's substantive, statistical, and operational aspects, and provides the technical details required for implementing the assessment. The student model (SM) specifies the variable(s) in terms of which we wish to characterize students. It may contain a single variable, representing an overall proficiency, or multiple variables, characterizing several aspects of knowledge or competences. Technically, the SM model can be presented by a possibly vector-valued parameter (usually denoted by θ), and a joint probability distribution p(θ). A student model then can be viewed as a mathematical structure containing variables that can take a range of possible values, and a joint probability distribution function quantifying relationships among these variables. Reasoning from observable behavior in task situations with given features, we can characterize the students' knowledge, skills, or proficiency which we are interested in making inferences about, and use probability distributions over a student's SM variables to express our belief about their values. Values of SM variables correspond to claims that can be made about students, for example, as to

their level of proficiency for getting correct answers in a domain of tasks, as in traditional testing, or as to the way they may be thinking about problems in the domain, as in Bao and Redish's research (2004). Later in this chapter, we discuss how alternative psychometric models express different conceptions of knowledge and learning in physics, including mixtures of qualitative knowledge states suggested by cognitive research but poorly addressed by traditional test analyses.

A task model (TM) in the CAF concerns substantive considerations about the features of tasks that are necessary to evoke evidence about SM variables. It embodies beliefs about the nature and structure of task situations, as they are important under the conception of knowledge that guides the assessment's design. With regard to work products (e.g., what the student says, does, or produces), the task model also specifies what student behaviors or productions will be observed as they provide clues about their knowledge, again as they are important under the conception of knowledge that guides the assessment's design. Therefore, for a particular task, the values of task model variables consist of information characterizing the situation with regard to its salient features and the kinds of performances that will be captured. In addition, the TM also describes features of tasks that are needed to inform the operational activities for particular assessment tasks (e.g., authoring, calibrating, presenting, and coordinating). Although many tasks can be created given a task model, the collection is constrained only to suit the needs of the assessment project. In this dissertation research, the tasks we use were designed in previous research and the data were collected by Bao and Redish to provide evidence about how college students learn and apply Newtonian mechanics in physics. We make explicit, in the framework of a task model, the connections between the Bao-Redish conception of

physics learning and the features of tasks in their test to reveal the mixtures of knowledge states they predict.

An evidence model (EM) in the CAF concerns reasoning from what we observe in a given task situation to update beliefs about SM variables. It contains two components, which connect students' work products to their knowledge and skill: the evaluation component and the measurement component. One can think of the evaluation component as "task scoring" since it describes rules for extracting evidence from individual performances, as values of observable variables. In other words, the evaluation component indicates how one identifies and evaluates (e.g., through rubrics) students' work or performance (what they say, do, or produce in a given task), and expresses salient aspects of them as values of observable variables (e.g., item or task scores). In comparison, one can view the measurement component as "test scoring," for it contains statistical models used to synthesize information or analyze data from observable variables across performances, in order to reflect belief about SM variables.

Technically speaking, the measurement component specifies models used to construct likelihood functions for SM variables (as induced by the values of the observed variables) and to estimate model parameters to obtain estimators for SM variables. Therefore, one can see that the measurement models in this context make connections between student models and task models. If tasks are well developed jointly with measurement model, one can take advantage of efficient statistical computing to use the complex model or estimation procedures (e.g., full Bayesian analysis) to support inferences in terms of preplanned and substantively important patterns in data. For the current study, while much work in the evaluation component has been done by Bao and

Redish, the little-known AR model is used to analyze data collected from Bao and Redish to examine college students' problem-solving in physics. Bayesian estimation procedures (Markov chain Monte Carlo techniques, MCMC) are used to estimate models' parameters and address their comparative fit to the data. The AR model and the basic idea of MCMC estimation procedure are introduced in later sections.

Finally, the last stage of ECD is the Operational Assessment. It concerns the operation of the implemented assessment based on the design generated in the previous stages (in particular, the CAF). Since this stage is not closely related to the current study (the data has already been collected), it is not discussed in detail here.

Science of Learning in Physics

In terms of ECD, we design a task with some features and hope to make an inference about the student variable(s) given the evidence in performances evoked by those features. By knowing how students learn and approach new physics concepts, one can design an appropriate task to examine their understandings, transitional stages, or misunderstandings, of those concepts. These three constructs (understanding, transitional stage, and misunderstanding) can be reflected by students' model use for problem-solving. They could use the correct model, mix correct with incorrect models, or simply use the incorrect model, as discussed later in this chapter. This perspective is naturally allied with improving instruction for abstract physics concepts, although it is an analytic method rather than a contribution to the substantive base that is the focus of the dissertation. By the way of background, then, this section provides some general ideas of development in the science of thinking and learning in terms of four perspectives (more details can be found on

pp. 57-110, National Research Council, 2001), focusing on a cognitive perspective. From this perspective, we emphasize the issues and concepts associated with physics learning, especially those relevant to Bao and Redish's point of view (2004).

Four perspectives. With regard to instruction and assessment, there are four perspectives that are particularly significant in terms of history of research and theory regarding the nature of the human mind (Greeno, Collins, & Resnick, 1996):

- Differential perspective.
- Behaviorist perspective.
- Cognitive perspective.
- Situative perspective.

These four approaches are not mutually exclusive; rather, each approach emphasizes different aspects of knowing and learning with different implications for what to be included in the task to reveal individuals' abilities or conceptual understanding and how to design and implement the task.

During the first few decades of the 20th century, researchers focused on how individuals differ in their general intelligence ability – the differential perspective. This approach assumes that individuals differ in their mental capacities and that these differences define stable mental traits or cognitive abilities (e.g., aspects of knowledge, skill, and intellectual competence) that can be measured. Individuals possessing different amount of these traits would show different levels (or patterns) of performance on tasks designed to reflect those abilities. Obviously, this approach is not aimed at studying how individuals process and store information in daily living or school learning.

Similarly, the behaviorist theories that became popular and dominated much of the research and theory on learning during the middle of the century also do not take cognitive processes into account in terms of learning. They view knowledge as the organized accumulation of stimulus-response associations that serve as the components of skills. These associations can be strengthened by reinforcement (e.g., rewards) or weakened by punishments, and this is what motivates individuals' learning. On one hand, many behavioral laws and principles have been derived from this perspective to promote learning or mediate behaviors (e.g., quit smoking). On the other hand, as mentioned earlier, the behaviorist approach ignores the underlying cognitive structures or processes that could have influences on external behaviors.

It was not until 1960s that the cognitive perspective started to emerge, due to advances in fields such as linguistics, computer science, and neuroscience. This approach helps to further study individual development and learning by using powerful technologies to observe behavior and infer cognitive functioning and underlying processes. It focuses on how people develop knowledge structures – how they process new information, how they integrate new information with the prior knowledge, and how they retrieve knowledge to solve problems. Compared with the differential and behaviorist perspectives that only focus on how much knowledge or skills individuals have, cognitive theories also emphasize what type of knowledge individuals have, and how it is organized. This latter aspect of acquiring knowledge becomes more important in current assessments. An important purpose that one would like to be able to address with assessment is not only to examine how much individuals know (e.g., how many items examinees answer correctly or incorrectly) but also to assess how, when, and whether they can apply what they know. To

be able to do this, however, requires more complex tasks (than traditional tests that target only on students' overall performance) to reveal information about how individuals respond to the questions through their cognitive processes, including, for example, reasoning strategies and evolving understanding over time.

The last perspective for studying human behaviors is the situative approach, also known as the sociocultural perspective. While the cognitive perspective focuses on individual thinking and learning, the situative perspective describes behavior at a different level of analysis. It studies how practical activity and contexts such as culture and ethnicity "mediate" individuals' behaviors. Based on this view, human behaviors or learning could be affected by the communities to which individuals belong – groups of people, large or small, who have the same goal or share some common interests (e.g., family and school). Therefore, in the context of assessment, some students may be better prepared than others to take a multiple-choice test because their parents or school teachers provide them more opportunities to practice. Some test items may favor a specific ethnic group because the questions ask are part of the culture in which those examinees live. It is not surprising that not every social activity is evenly distributed among the population of test takers. Although some statistical techniques (e.g., differential item functioning in IRT) have been developed to detect such bias among examinees, most current testing practices have been found to be not a good match with the situative standpoint. The reason is that the situations presented in most tests are not well connected with the specific contexts in which people typically use the knowledge or skills being tested.

In the current study, we take the cognitive approach to study how students solve physics tasks. In particular, we are interested in examining how individuals organize information or knowledge that is processed. Knowing this will help us understand how people answer questions or solve problems. This, in turn, can help to improve instruction. Below we present key ideas derived mainly from a cognitive perspective to understand students' learning in physics from Bao and Redish (2004). We also review selected studies in physics to show how students' prior knowledge affects their learning new concepts, a concern also shared by Bao and Redish.

Bao and Redish's view about students' learning in physics. Bao and Redish (2004) presented their view about students' learning based on a synthesis of three kinds of scientific research: ecological (phenomenological observations of normal behavior), psychological (studies of cognitive structures), and neurological (studies of the structure and functioning of the brain). Although the final model of learning has not yet been determined, they considered the following three elements as important factors in studying students' problem solving in physics – memory (in particular, long-term memory, focusing on its characteristics), context dependence, and the structure in long-term memory.

Memory is one of critical issues for both teaching and learning. Bao and Redish focused on long-term memory because they were interested in evaluating the success of instruction and this involves how students utilize their long-term memory. Three characteristics of long-term memory were particularly related to their study:

1. Long-term memory can exist in (at least) 3 stages of activation: inactive, primed (ready for use), and active (immediately accessible).
2. Memory is associative and productive. The elements stored in the long-term memory are associated, so activating one element leads, with some probability, to the activation of associated elements. Activation often consists of data receiving, reasoning, and mapping the memory elements onto input structures.
3. Activation and association are context dependent, both external and internal (other activated elements).

They also believe that how students respond to a physics question depends on the interactions between students and the historically and culturally constituted contexts. In fact, this reflects the situative perspective as discussed earlier. The evidence of context dependence can be found from the physics education literature. For example, it can be shown that students respond differently on two formally equivalent questions if they are stated in two different scenarios (i.e., one is phrased in physics terms using a laboratory example; the other is phrased in common speech using everyday experience). The majority of students tended to answer the physics-like problem correctly, while only a half of students answered correctly on the everyday problem (Steinberg & Sabella, 1997). This is particularly true when students are just beginning to learn new material. It could be due to the fact that students have not yet mastered the new concept and do not know how to apply the knowledge or skills they have learned to the situation they encounter or to the question they are asked.

Five structures in long-term memory that are particularly relevant for the understanding of physical phenomena and for the study of physics were identified by Bao and Redish from the literature of science learning: *reasoning primitives*, *facets*, *schemas*, *mental models*, and *physical models*. Note that these structures describe how people organize information or knowledge within each structure. They differ in the way students use them for reasoning or problem-solving, as discussed below.

A *primitive*, in Bao and Redish's definition, concerns the finest-grained cognitive element. A reasoning primitive is what people use to describe why things work the way they do, and the typical response based on a primitive is "That's just the way things work." They cannot give a reason why things happen. diSessa (1993) refers to such statements as " phenomenological primitives" or "p-prims". The term *facet* refers to the mapping of a reasoning primitive into a physical situation.

The next terms in Bao and Redish's hierarchy are *schema* and *mental model*. They are broadly defined, and play a critical role in their study in understanding students' responses. A schema refers to a cognitive element or a set of cognitive elements that are activated together in response to a stimulus or situation presented to the student, while a "robust and reasonably coherent" schema refers to a mental model. Finally, a *physical model* is a type of mental model commonly used by a certain population to describe the characteristics and properties of a specific physical object or set of objects. Based on this definition, people may view the same object or system differently. For example, they could describe it in terms of macroscopic point of view or view it from a microscopic perspective.

How do these cognitive elements described above (i.e., context-dependence and the structures in long-term memory, especially schemas and mental models) operate in students' learning? Bao and Redish (2004) carried out research in the context of sets of "expert-equivalent questions", a sequence of questions or situations in which an expert would use a consistent mental model, but, based on the literature in physics learning, a particular student can use any of a variety of mental models (e.g., McCloskey, op cit.; Viennot, op cit.). They hypothesized that a student's response not only depends on his/her educational history (i.e., the previous experience or the preexisting knowledge about a specific concept) as will be discussed below, but also on the student's mental model state at the particular instant triggered by the question presented to him/her. The latter proposal suggests that in at least some cases a student will be in a mixed model state, indicating that he/she can be thought of as simultaneously occupying a number of distinct model states[1], and which state would be invoked depends on the features of a particular question. They also suggested that the appropriate way of analyzing this situation (mixture-within-persons) is to study the student's responses using a probabilistic mathematical model. Their method is described in more details later in this chapter.

As mentioned above, students' previous experience or prior knowledge may have impacts on students' physics learning. In his study, diSessa (1982) found that regardless of

[1] For examples, Newtonian way of thinking, a belief consistent with Newton's laws; "impetus theory" belief, stating that a certain force keeps acting on a moving object until the force is diminished by other forces such as those from air or gravity; and an Aristotelian way of thinking, an expectation that objects simply move in the direction you push them without considering the combination of forces or other naïve beliefs about forces.

previous physics learning in high school and in the freshman year, a particular college student used strategies quite similar to those of elementary students in that they both started to approach this task with Aristotelian conceptions of physics, indicating that the college student could not make the connection between what she had learned in physics (e.g., vector addition and conservation of momentum) and how to complete the task. It was further suggested that previous knowledge would play an important role in understanding the new physics concepts (students were trying to understand the new concepts using previous experience) and in applying them to the novel situations. In other words, when a new task is given to physics novices, they tend to use their preexisting knowledge or experience to solve for it. That is why the Aristotelian-type strategies were immediately adopted by elementary students or even the college student for problem-solving. The Aristotelian expectation is closer to common sense and everyday "intuitive" manipulation of the world, but is contradictory to Newton's laws.

In their reviews on misconceptions of the concept of force, light, heat, and electricity, Reiner, Slotta, Chi and Resnick (2000) provided evidence that naive conceptions often reflect an underlying commitment to preexisting knowledge of material objects or substances. Using force as an example, in physics, force is seen "as a process of interaction involving two or more material objects in which these objects are sped up, slowed down, or redirected" (p. 10). However, physics novices do not conceive of force as a process of interaction between two material objects. Rather, they think of force as either some act of the object itself or a property of a material object (e.g., novices tend to explain gravity by assuming an innate, inexhaustible internal property called weight, or even explain the fact that an object will fall if dropped because of the contact of air pressure).

Understanding how prior knowledge affects students' learning would help us to explain, for example, why students tend to use "impetus theory" or the Aristotelian approach for solving Newtonian mechanics problems even after instruction. Our analyses in Chapter IV show some examples of such analyses.

In the current study, sets of tasks identified by Bao and Redish were used to examine whether college students apply Newtonian thinking in conceptual problem-solving. Based on what researchers have suggested about physics learning, as reviewed above, we expect that students may approach questions in different ways depending on the degree to which they have mastered the concept (naïve physics learners may use Aristotelian theories or other preexisting knowledge to answer questions, while those who understand Newtonian mechanics will tend to use Newtonian way of thinking to respond to given tasks) and on the features of the questions. Therefore, our goal here is to use appropriate psychometric models to see if we can provide evidence to support what we believe about how students apply abstract physics concepts. The AR model, along with other IRT models (in particular, the 3PL) and LC models are reviewed below.

Measurement Models

Three psychometric models are discussed here. Each model represents a distinct perspective in terms of students' learning, and they are different from the traditional test analysis (e.g., based on the total score or the percentage of answering items correctly). Our main interest lies in the AR model, because this is the one closest to the psychological/substantive model that motivated the items. The other two models – an IRT model and a LC model – are included partly to contrast them with the AR model, but even

more to be able to discuss the interplay among substantive perspectives, task design, measurement models, and inferences about students. However, as mentioned in the introduction, our main purpose in the current study is not to choose the best fitting model but to analyze data from a model that embodies the kinds of patterns that its psychological grounding entails. Therefore, only the AR model is used in the data analysis.

Note that the AR model is a Rasch-type IRT model. It can be considered as a multivariate (or multidimensional) IRT model since, as discussed in a later subsection, more than one parameter is associated with each person, whereas conventional IRT models deal with a unidimensional latent trait. In addition, the AR model is appropriate for polytomous data (i.e., data with multiple response categories) while many conventional IRT models deal with binary data (i.e., data is coded, for example, as correct or incorrect).

In general, IRT assumes that, in testing situation, examinee performance on a test can be explained by his/her underlying latent variable (e.g., ability). In fact, it can be further extended to any situation in which a paper and pencil test or a questionnaire is used to measure educational or psychological constructs. The primary use of IRT is for modeling examinees' propensities to give higher quality responses – right rather than wrong multiple choice responses, for example, or well-constructed rather than poorly-constructed essays. An IRT model is then developed to specify the relationship between the "observable" (e.g., item scores) and the "unobservable" quantities (e.g., abilities) by a mathematical probabilistic function. It yields the probability of a correct response to an item (or more broadly, the probability of a response in a particular response category) given the examinee's position on the continuum (e.g., the ability or propensity level), the item's position on the continuum (e.g., the item difficulty parameter), and other

possible item parameters (e.g., the item discrimination parameter and/or the guessing parameter). In brief, an IRT model simultaneously takes into account the individual's ability (or other latent traits) and the item characteristics for analyzing responses. Interested readers can refer to Hambleton and Swaminathan (1985) for a more thorough discussion on conventional IRT models (mainly dealing with unidimensional latent traits for binary data) and van der Linden and Hambleton (1997) for modern IRT models (extending to items with polytomous response formats and/or multidimensional latent traits).

One of common and essential assumptions for IRT and other measurement models (e.g., LC models) is the assumption of local independence (LI), or conditional independence (CI) of item responses given item and person parameters. LI states that an examinee's responses to different items in a given test are statistically independent, given the item and person parameters in the model. In other words, an examinee's response on one item must not affect (in any ways) his or her performance to any other items in the test, above and beyond the relationships that are accounted for by those parameters. (In the case of typical IRT models, the student parameters correspond to ability in the domain of tasks; as seen below, the student parameters correspond to tendencies to employ different problem-solving approaches.) Obviously, it is not always the case in practice, so the issue becomes to what extents and in what ways the assumption of LI is violated. There are standard ways to check for when and how it holds. For longer tests with enough data per examinee, say twenty items or more, tests such as item-fit indices can be carried out to examine the assumption of LI (e.g., Q_2 test proposed by Van den Wollenberg, 1982; Q_3 test proposed by Yen, 1984, & 1993; R_2 test developed by Glas & Verhelst, 1995; and S_3 and

LM test proposed by Glas & Falcón, 2003). Those statistics may not be appropriate to be used for tests with a few items (this is the case for the current study since the data only contains no more than 8 items). However, we can check the fit of the AR model that assumes LI over various contexts by, for example, comparing item parameter estimates before and after instruction. This approach to checking a particular LI violation is discussed in the next chapter.

The following subsections review the form of measurement models (again focusing on the AR model), and compare the other two models with the AR model in terms of the substantive interpretation of the student model they would imply as they relate to the assessments Bao and Redish studied.

The Andersen/Rasch (AR) multivariate measurement model. The AR psychometric model is consistent with the above-reviewed literature concerning students' cognitive process in physics learning for the measurement paradigm instantiated in assessments such as the FCI and FCME. Based on the literature, a student's mental model state could consist of distinct competing physics models, and which one would be used for problem-solving depends on the features of the item presented to him/her. The AR model is an appropriate one from psychometrics since it was developed to model students in terms of propensities toward characteristic types of response which in this case will be conceptions of the domain (it can be viewed as a "mixture-within-persons" approach, as further explained below), rather than in terms of expected correctness. The analysis based on the AR model addresses the same kinds of student/task interactions as the *Model Analysis* methodology that Bao and Redish (2004) developed in their research to study students' physics learning.

The AR model states that at a given point in time, a person is seen as having propensities to answer in accordance with any of the conceptions. Tasks are also parameterized in terms of their tendency to provoke different conceptions as well. Given physics tasks (e.g., a multiple-choice test with *j* items, each item having *m* choices that are each associated with a particular way of thinking about situations in the targeted domain) administered to *N* examinees, the idea of AR model can be presented as follows.

In terms of items, each item can be characterized by a vector value containing *m* elements, with each value of the element corresponding to a location on a continuum for a certain property in physics in a sense that can be described as follows. For example, the first choice may represent the Newtonian approach (those who pick this choice have behaved in a way consistent with a Newtonian strategy for problem-solving), the second choice may represent the strategy using "impetus theory", the third choice may reflect an Aristotelian belief, and so forth. In other words, the item parameter is a vector-valued parameter which contains *m* elements, and larger values for an element indicate a greater tendency for that item to elicit responses in line with the corresponding problem-solving approach. In particular, the choice with the highest value indicates that that way of thinking is more common on this item, all other things being equal. In line with the science-learning research noted above, particular features of given items can tend to evoke particular misconceptions.

For persons, similar to items, each examinee is characterized in terms of a vector-valued parameter that also contains *m* elements, with each element representing the associated propensity level on the continuum. If, for instance, person A has the greatest propensity level on the Newtonian approach (the first choice on the above example), this

indicates that person A tends to respond in accordance with Newtonian strategies for problem-solving. Furthermore, if the task is designed to examine whether or not students have mastered Newton's third law, then we can make an inference about person A saying that he/she probably understands how to apply Newton's third law in this situation.

Let X_{ij}; $i = 1, \ldots, n, j = 1, \ldots, k$ be independent random variables (i is the index for examinees while j is the index for items) and further assume that there are m discrete choices for each item (so X_{ij} can be any integer between 1 and m). The m options are associated with kinds of response that are the same across all items, with respect to strategy, perspective, style, conception, or some other way of partitioning responses in the domain. In our example above, category 1 responses are consistent with a Newtonian approach, category 2 with an "impetus theory" approach, and category 3 with a naïve or Aristotelian approach, which might be called a null approach. The formal AR model can be written as:

$$P(X_{ij} = p) = \exp(\theta_{ip} + \beta_{jp}) / \sum_{p=1}^{m} \exp(\theta_{ip} + \beta_{jp}), \qquad (1)$$

where

p is an integer between 1 and m;

θ_{ip} is the pth element in the person i's vector-valued parameter; and

β_{jp} is the pth element in the item j's vector-valued parameter.

Again, note that there are m probabilities for each examinee on a given item, representing the probability of choosing any particular choice for that person on that item.

The graphical representation of the AR model can be viewed as item response surfaces (IRSs) (not the item response curve, since the model is multivariate). For a given item, there are more than one IRSs under the AR model, depending on the number of response categories (i.e., m) – one IRS for each response category. In addition, each IRS is multidimensional since there is more than one parameter for a given person. If m equals 3 (as the example above) and a vector-valued item parameter is given, one can have *three* three-dimensional IRSs as functions of the values of person parameters θ_{i1}, θ_{i2}, and the corresponding probability as dimensions of the graph. (Note that with three response categories there are only two unique person parameters since to identify the model the constraint needs to be imposed that $\theta_{i3} = -(\theta_{i1} + \theta_{i2})$ – this is discussed below. Thus θ_{i3} can be omitted when plotting IRSs.)

Figures 1-3 represent the IRSs given that an item's vector-valued parameter (β_j) is (2.5, .5, -3.0), an item with a greater tendency to evoke the response category 1. From Figure 1 (IRS for the response category 1), one can see that as the value of $\theta_{.1}$ increases (while the others remain the same), the probability of choosing the response category 1 increases as well. By the time that $\theta_{.1}$ equals 4, the probability is approaching 1. On the other hand, the probability is not much changed as the value of $\theta_{.2}$ goes up. A similar interpretation holds for Figure 2. For Figure 3, the probability for selecting the response category 3 reaches the peak when both θ_{i1} and θ_{i2} decrease – so θ_{i3} increases. However, the probability will not approach 1 because the given item has least tendency to provoke response category 3.

Figure 1.
Item Response Surface for the Response Category 1

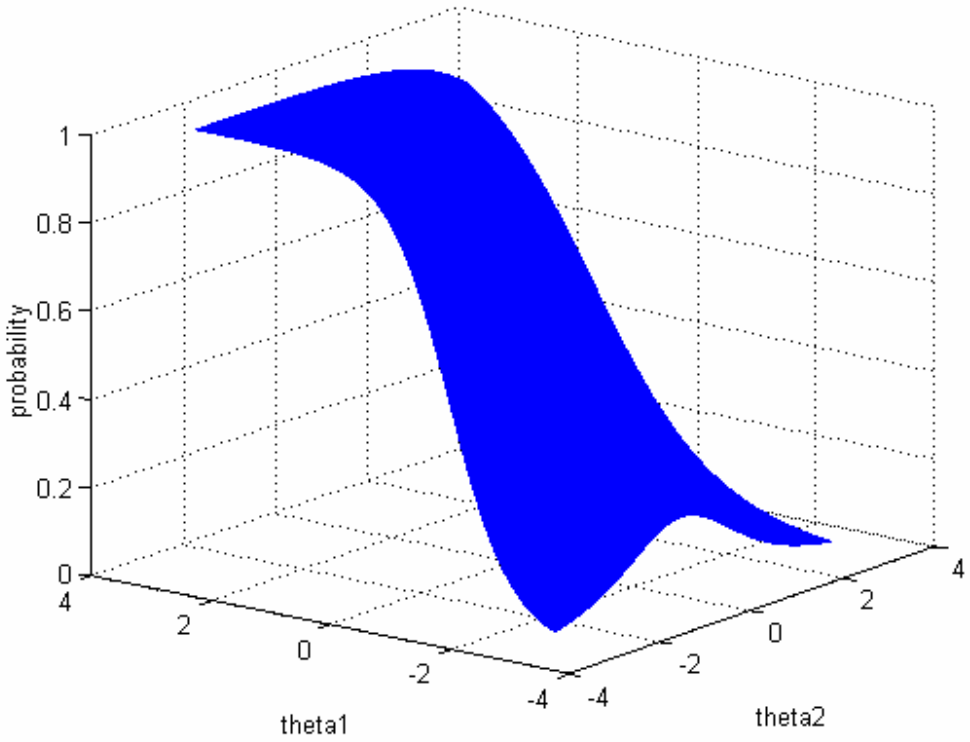


Figure 2.
Item Response Surface for the Response Category 2

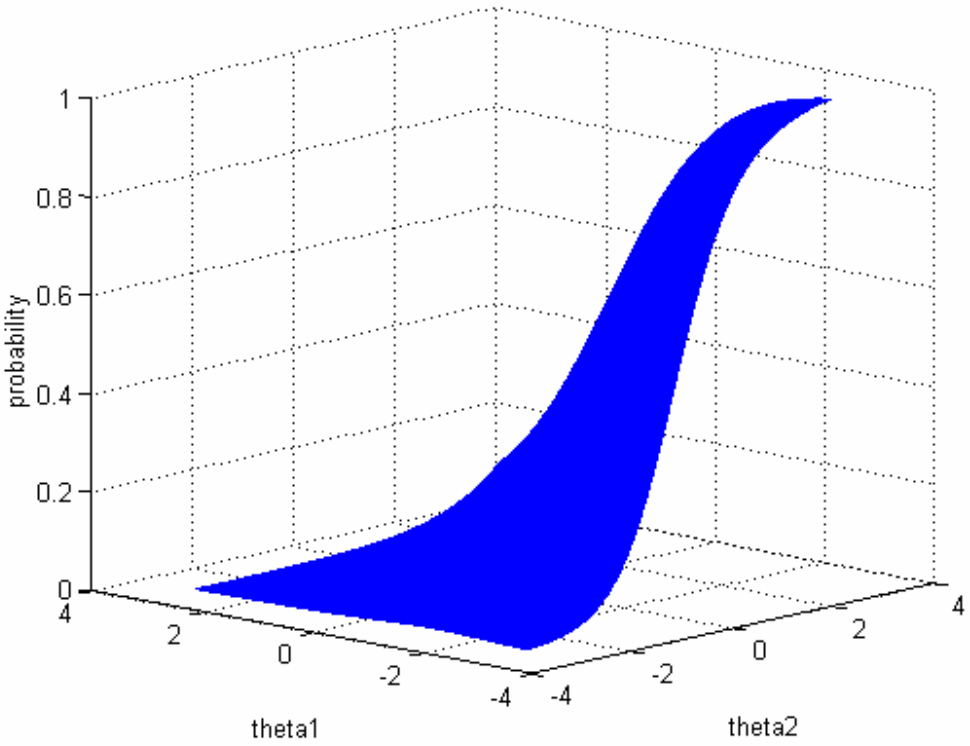
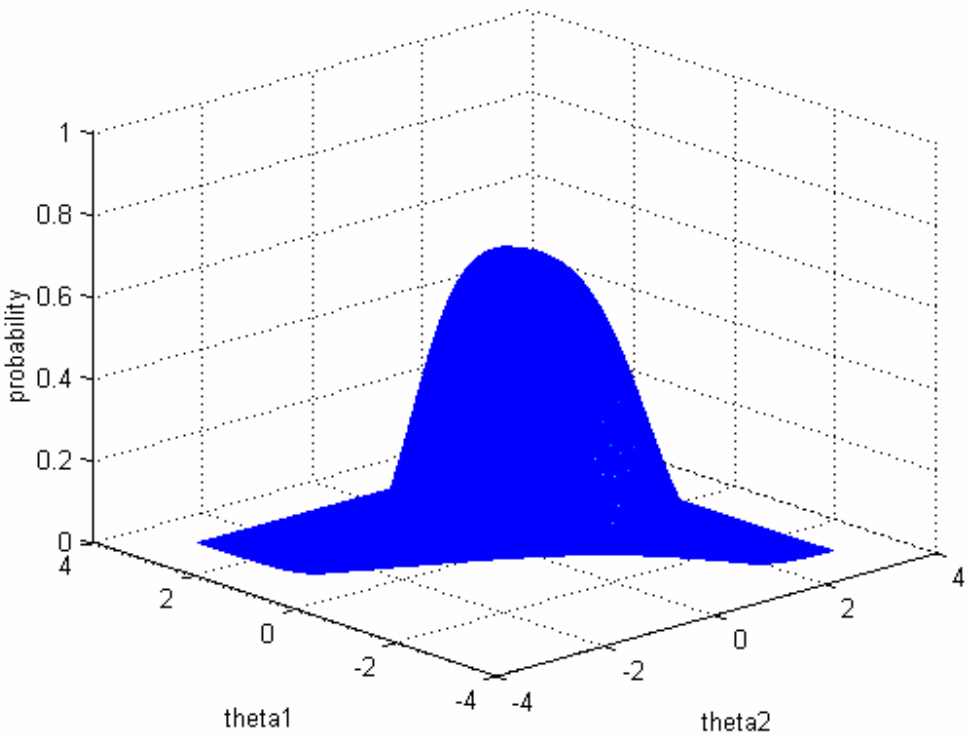


Figure 3.
Item Response Surface for the Response Category 3



As in other IRT models, adding a constant to all item and person parameters will result in the same probability. In regard to removing this indeterminacy of the model, one way the scale can be fixed by centering parameters for each item and person around zero, i.e., $\beta_{j3} = -(\beta_{j1} + \beta_{j2})$ for item *j,* and $\theta_{i3} = -(\theta_{i1} + \theta_{i2})$ for person *i*. In addition, setting normal priors with fixed means for item and person parameters, a common practice in Bayesian IRT estimation procedures, ensures the identification of the AR model. This is the method used in the *BUGS* analyses discussed in Chapter IV.

In order to examine if the AR model is identifiable (i.e., the parameter estimates are unique given a data set), an equivalent set of constraints will be utilized within the framework of the multidimensional random coefficients multinomial logit model (MRCMLM) developed by Adams, Wilson and Wang (1997). The MRCMLM can be used to represent a wide class of Rasch models as special cases, including: (1) unidimensional models such as the linear logistic model (Fischer, 1973 & 1995), the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982), and the ordered partition model (Wilson, 1992); and (2) multidimensional models such as the multicomponent latent trait model (Whitely, 1980) and the multidimensional Rasch model for learning and change (Embretson, 1991). We can show that the AR model is also a special case of MRCMLM using a proper parameterization. Then we can apply the conditions given and proven by Volodin and Adams (1995) that are necessary and sufficient for the identification of MRCMLM to the case for the AR model. These conditions, developed for use under maximum likelihood estimation, express conditions for suitably constrained sets of item parameters.

The formal MRCMLM can be represented by:

$$P(X_{jk} = 1; A, B, \beta \mid \theta) = \frac{\exp(b_{jk}\theta + a'_{jk}\beta)}{\sum_{k=1}^{K_j} \exp(b_{jk}\theta + a'_{jk}\beta)}, \quad (2)$$

where: $X_{jk} = 1$ indicates a response is in category k for item j and $X_{jk} = 0$ indicates it is not; β is a column vector containing P item parameters being estimated; A, the design matrix of the test, is used to impose a linear relationship among item parameters and is defined by design vectors a_{jk} (for $j = 1,\ldots,J$; and $k = 1,\ldots,K_j$), so that if one defines $K = \sum_{j=1}^{J} K_j$, A is the $K \times P$ matrix of the form $A = (a_{11}, a_{12}, \ldots, a_{1k_1}, a_{21}, \ldots, a_{2k_2}, \ldots, a_{Jk_J})$; θ is a $D \times 1$ column vector containing D person parameters being estimated for each person; and B, the scoring matrix of the test, is formed by $(K_j - 1) \times D$ submatrices B_j that is the catenation of $(K_j - 1) \times 1$ column vectors b_{jk} giving the performance level of an observed response in category k of item i. If a response category for a particular item does not relate to a particular latent dimension, then the score on that latent dimension is set to zero.

Consider a test with five items and three response categories for each item (corresponding to the Newtonian, "impetus theory", and Aristotelian approaches, respectively, as the example used earlier). By using appropriate constraints (and using the notations for the AR model) and choices of A and B matrices, it can be shown that the AR model is a special case of MRCMLM. First, we impose $\beta_{j1} = 0$ for every item j (i.e., every item's parameter for Newtonian approach is set to 0) and $\theta_{i1} = 0$ for every person i (i.e., every person's parameter for Newtonian approach is also set to zero). These two constraints are necessary to remove the indeterminacy of the AR model, and help us

identify both A and B matrices needed to construct the AR model in the context of MRCMLM. (The "set first component for all items and all persons" constraints are equivalent to the "sum-to-zero within items and within persons" constraints described in the previous section and used in the parameter estimation in Chapter IV, but are expressed in the form also used for identification conditions that will be discussed below.)

The matrix B will take the following form:

$$B = \begin{bmatrix} 10 \\ 01 \\ 10 \\ 01 \\ 10 \\ 01 \\ 10 \\ 01 \\ 10 \\ 01 \end{bmatrix}, \tag{3}$$

where the first and second column represent the "impetus theory" and Aristotelian dimension, respectively; and the first two rows are for item 1, the next two rows are for item 2, and so forth – the first row for a given item is for the response category 2 and the second row is for the response category 3. Notice that there is no Newtonian dimension due to the constraint we impose above regarding θ. Similarly, there is no parameter associated with the response category 1 for any item.

The matrix A will take the following form:

$$A = \begin{bmatrix} 1000000000 \\ 0100000000 \\ 0010000000 \\ 0001000000 \\ 0000100000 \\ 0000010000 \\ 0000001000 \\ 0000000100 \\ 0000000010 \\ 0000000001 \end{bmatrix}, \quad (4)$$

where the first two columns are for the "impetus theory" and Aristotelian parameters of item 1, the next two columns are for item 2, and so on. Altogether the ten rows are the "impetus theory" and Aristotelian responses for all five items (two response categories for each item).

However, one more constraint in this type of parameterization is necessary to further remove the indeterminacy of the AR model (Andersen, 1973, pp. 144-145) due to the linear relationship between θ and β in Equation (1) – the probabilities remain the same for each response category p when the same constant is added to θ_{ip} and subtracted from β_{jp}. One possibility is to set both β_{12} and β_{13} equal to 0, that is, β_{1p} is 0 for any p (after combining the first two constraints) – using Item 1 as a baseline. (One could arbitrarily choose any item and set all of its parameters zero.) Under this parameterization, the person parameters for the two dimensions being estimated need no further restrictions. Notice again that this style of parameterization is different from the one mentioned previously (i.e., centering parameters for each item and person around zero, then incorporating priors to identify the centers of the θ distributions) but they accomplish the same end.

Then the matrix A is reduced to the following form:

$$A = \begin{bmatrix} 00000000 \\ 00000000 \\ 10000000 \\ 01000000 \\ 00100000 \\ 00010000 \\ 00001000 \\ 00000100 \\ 00000010 \\ 00000001 \end{bmatrix}. \qquad (5)$$

The eight columns in (5) are for the "impetus theory" and Aristotelian parameters of items 2 through 5 only (two columns for each item), and the ten rows are the "impetus theory" and Aristotelian responses for all five items.

For the normal person-parameter case, Volodin and Adams (1995) showed that the following are necessary and sufficient conditions for the identification of MRCMLM:

Proposition 1. If D is the number of latent dimensions, P is the length of the parameter vector, β, $K_j + 1$ is the number of response categories for item j, and $K = \sum_{j=1}^{J} K_i$, then MRCMLM can only be identified if $P + D \leq K$.

Proposition 2. If D is the number of latent dimensions, P is the length of the parameter vector, β, then MRCMLM can only be identified if rank(A) = P, rank(B) = D, and rank([B ||A]) = $P + D$. (Note: [B || A] refers to the horizontal concatenation)

Proposition 3. If D is the number of latent dimensions, P is the length of the parameter vector, β, $K_j + 1$ is the number of response categories for item j, and $K = \sum_{j=1}^{J} K_i$, then MRCMLM can only be identified if and only if $rank([B \parallel A]) = P + D \leq K$.

Using the matrices of B and A, shown in (3) and (5), as the parameterization of MRCMLM for the AR model, and according to the definitions given by those propositions above, we can see that:

1) $D = 2$, since only two dimensions are being estimated (i.e., "impetus theory" and Aristotelian approach);

2) $P = 8$, since only eight item parameters are being estimated (2 parameters each for items 2 through 5 only);

3) $K = 10$, since there are three response categories for each item;

4) $rank(A) = 8$ based on (5);

5) $rank(B) = 2$ based on (3); and

6) $rank([B \parallel A]) = 10$ since the matrix of $[B \parallel A]$ takes the following form:

$$[B \parallel A] = \begin{bmatrix} 1000000000 \\ 0100000000 \\ 1010000000 \\ 0101000000 \\ 1000100000 \\ 0100010000 \\ 1000001000 \\ 0100000100 \\ 1000000010 \\ 0100000001 \end{bmatrix}. \quad (6)$$

Therefore, Proposition 1 is true for our case since $P + D = 8 + 2 = 10 \leq K$. Proposition 2 is also true since $rank(A) = 8 = P$, $rank(B) = 2 = D$, and $rank([B \parallel A]) = P + D = 10$. Finally, Proposition 3, following the first two propositions, is also valid here: $rank([B \parallel A]) = P + D = 10 \leq K(= 10)$.

Thus the AR model is identifiable. One can use the same approach to examine if the AR model is identified for a data set other than five items or three response categories by using the same style of setting constraints and checking the propositions.

It may be noted that since the AR model belongs to the Rasch family, sufficient statistics exist for both items and persons. For person *i*, the minimal sufficient statistic is the number of X_{ij}'s ($j = 1,...,k$) with observed value *p*, and the person's score can be determined by the weight associated with each response category (so the sufficient statistic for person *i* would be the sum of weighted response categories). Similarly, for item *j*, the minimal sufficient statistic is the number of X_{ij}'s ($i = 1,\ldots,n$) with observed value *p*.

The three-parameter logistic (3-PL) IRT model. For binary data, the most general and commonly used IRT model is the 3-PL model proposed by Birnbaum in the late 1950s. The 3-PL model takes the following form:

$$P_j(\theta) = c_j + (1 - c_j)\frac{\exp a_j(\theta - b_j)}{1 + \exp a_j(\theta - b_j)}, \qquad (7)$$

where $P_j(\theta)$ is the probability that an examinee with ability level θ answers item *j* correctly; b_j is the difficulty parameter for the item *j*; a_j is the discrimination parameter for the item *j*; and c_j is the guessing parameter for the item *j*.

The 3-PL model is adequate for a multiple-choice with a single response test in which examinees may obtain answers by guessing. In addition, it is assumed that θ is unidimensional.

Latent class (LC) models. Unlike IRT in which the underlying latent variable (θ) is assumed to be continuous, the latent variable assumed in a LC model is categorical since its purpose is to predict memberships (or classes). For a set of V dichotomous variables, there are 2^V different response patterns that can be observed although some of them may not be seen in real data as mentioned above. In general, the responses for a sample of N cases can be summarized in a frequency table which shows the 2^V response vectors along with the number of cases associated with each pattern. This table is then to be analyzed. The Equations (8), (9) and (10) below represent the general form of latent class analysis (van der Heijden, Dessens, and Bockenholt, 1996):

$$\pi_{uvwx} = \pi_x \pi_{u|x} \pi_{v|x} \pi_{w|x}, \quad \forall u, v, w, x, \tag{8}$$

with restrictions

$$\sum_{x=1}^{X} \pi_x = 1 \quad and \quad \sum_{u=1}^{U} \pi_{u|x} = \sum_{v=1}^{V} \pi_{v|x} = \sum_{w=1}^{W} \pi_{w|x} = 1, \quad \forall x, \tag{9}$$

where

π_x is the probability of falling into latent class x; and

$\pi_{u|x}$, $\pi_{v|x}$, and $\pi_{w|x}$ are the conditional probabilities of falling into levels u, v, and w, respectively, given x.

The unobservable probabilities are related to the expected probabilities π_{uvw} by Equation (10):

$$\pi_{uvw} = \sum_{x=1}^{X} \pi_{uvwx} \ . \qquad (10)$$

That is, the overall, or unconditional probability for a response vector (π_{uvw}) is the sum of each conditional probability for the response vector weighted by the corresponding latent class proportion.

Comparisons among the three measurement models with regard to the Bao-Redish assessment. IRT models are the most familiar item-level test theory models that concern students' propensities to do well on tasks in a defined domain; e.g., right answers rather than wrong answers, high ratings rather than low ratings on essays. The 3-PL model, applied to the Bao-Redish data, would therefore characterize students simply as to their propensity to make correct (Newtonian) answers to the items, or an overall proficiency level in students' learning. This is the kind of inference one can make about a student model through the 3-PL model. As discussed above, the 3PL would also characterize items as to their operational properties such as difficulty, discrimination, and/or guessing parameters.

The 3-PL model would be appropriate for the Bao-Redish data sets which were collected from the tests consisting of multiple-choice items if the objective of the analyst were simply to characterize students in terms of their propensity to make correct responses. However, an analysis based on the 3-PL model would not be able to probe the details of the processes students may have used for problem-solving. The 3PL embodies a certain way of thinking about students and items – which characteristics are important, and how those

characteristics are, in probability, reflected in students' performances. It is a student's tendency to produce right rather than wrong answers. This is the major difference between the AR model and the 3-PL model.[2]

Other IRT models have been developed to incorporate knowledge from cognitive science of learning into psychometric analyses (e.g., Embretson, 1995; Samejima, 1995; Wilson, op cit.). They are not conformable with current research in that the cognitive process hypothesized and parameterized in those models is not consistent with Bao and Redish's model for how students learn physics, and those models are not coded for the data sets used in the current study (i.e., response categories represent various strategies possibly used by students for problem-solving, and the responses are not necessarily involved in a systematic relationship to cognitive approaches).

The student model that accords with a LC model presumes that each student is a member of a class, associated with distinct probabilities for responses of different types. Applied to the Bao-Redish assessment, one might posit that each student is associated with a certain conception (or misconception) of Newtonian mechanics problems, and tends to respond in that way. These might be called "pure states" of understanding and expected

[2] It is possible that some polytomous IRT models (e.g., Masters' partial credit model – Masters, op cit.; or Muraki's generalized partial credit model – Muraki, 1992) could be used to obtain students' overall proficiency levels. These models are not widely used as the 3-PL model in the current educational measurement. The more important point is that recoding students' responses based on their level of partial knowledge as opposed to their tendencies to use different approaches would not be coherent with the task design in this study. These models are not considered here.

responses. There are occasional inconsistencies in responses, however, so that answers corresponding to other conceptions would occur with probabilities to be estimated.

In this sense, an LC model could be posited which would be similar to the AR model in that it can be used to model students in terms of conceptions/misconceptions of the domain rather than in terms of expected correctness (as in the 3-PL model). However, the LC differs in that a given student may not be in an arbitrary "mixture state" as in the AR model, but is instead modeled as in a "pure state" – that is, a student uses a consistent theory or model to respond, although there are some probabilities of responses of other strategies. Therefore, the latent class analysis (LCA) provides a different perspective in understanding students' learning – one that is farther than the AR model and even the 3-PL model in terms of investigating individual student's learning. It should be noted that some mixture LC/IRT models exist in the literature (e.g., Mislevy & Verhelst, 1990; Rost, 1991); however, they too are not considered here because those models concern propensity toward correct answers under different strategy uses. Neither the modeling of strategies nor the form of the data required for those analyses are consistent with the framework of the Boa-Redish analyses. The present aim is not to explore a given data set with a compendium of models, but rather to fit a model (the AR model here) most aligned with the intention of the assessment and to draw out the cognitive basis of that model.

The AR model matches up with the literature in cognitive science of physics learning in the area of focus, and is fully compatible with the Bao-Redish study. This is the main point in the current study – i.e., to show an example of using a psychometric model that is consistent with the studies in science of learning, and to make inferences about students by examining their responses to tasks designed under the same conception of

learning. The 3-PL model, which shows the degree to which students have learned, and LC models, which reveals students' learning in terms of class membership, are interesting ones and provide different perspectives in students' learning, but they are less well aligned with current development in the cognitive science of learning in physics in the area Bao and Redish addressed. These models are discussed here to contrast with the AR model in terms of the way they model students' responses, and the nature of inferences they can support regarding students' learning.

MCMC Estimation

Both item and person parameters under the AR model will be estimated by MCMC sampling-based methods using a Bayesian approach. Compared with other estimation procedures (e.g., joint maximum likelihood estimation, JMLE, and marginal maximum likelihood estimation, MMLE, which has become standard IRT methodology in practice), one advantage of using a Bayesian approach is that the estimation is direct given that the priors are specified in advance; therefore, it requires a much smaller sample size than maximum likelihood estimation (MLE) procedures to yield stable estimates. Also, no artificial constraints need to be imposed on the parameter space as with MLE, since outward drifts of the estimates are naturally and effectively controlled by the priors. Swaminathan and Gifford's simulation study (1986) on the 3-PL model showed that the Bayesian estimates stay in the parameter space. Furthermore, the Bayesian estimates show a closer relationship to the true values than JML estimates. Although from a frequentist perspective, Bayesian estimates are biased toward the mean of the prior distribution (i.e., exhibits shrinkage), the use of priors is also what keeps the parameter estimates (especially

the *a* and *c* parameters in the 3-PL model) in the admissible parameter space. Another advantage of using Bayesian estimation over MLE (especially MMLE) is that uncertainty in item parameter estimates are easily incorporated into examinee inferences, and vice versa (e.g., Kim, 2001; Patz & Junker, 1999a; Tsutakawa & Johnson, 1990). Finally, Bayesian estimation provides solutions for those examinees with perfect or zero scores. JMLE fails unless those examinees are removed prior to estimation.

MCMC techniques have been recently applied to estimate parameters for latent variable measurement models, especially IRT models. For example, Albert (1992) applied a Gibbs sampling method to estimate item parameters under the two-parameter normal ogive model for a 33-item Mathematics Placement Test administered to 100 examinees. He then compared the estimates with those derived from MLE using EM algorithm based on a normal approximation. It was found that in terms of item difficulty parameters, these two estimation procedures yielded similar results; for discrimination parameters, the estimates based on Gibbs sampler tended to be larger than those resulting from MLE/EM, indicating that the marginal posterior distributions exhibited right skewness. By examining the standard error of estimates, he further suggested that the normal approximation to the posterior of the item parameters based on the mode and information matrix (used to compute the EM standard errors) might be a poor approximation to the exact posterior distribution.

Since these findings are only based on a single test, Baker (1998) conducted a simulation study to further investigate the item parameter recovery characteristics of Albert's Gibbs sampling method by comparing with those obtained from BILOG (Mislevy & Bock, 1989) in which MMLE/EM was implemented. He found that for a data set with

50 items and 500 examinees, the item parameter recovery of both estimation procedures was excellent. For a fewer number of items and examinees BILOG tended to be superior to the Gibbs sampling although the differences were small. As suggested by Baker, this could be due to the program Albert developed to implement Gibbs sampling. A more highly developed "production" version of the program might produce better results. Albert's Gibbs sampling procedure was also used by Fox & Glas (2001) to estimate parameters for a multilevel IRT model. In addition, Beguin and Glas (2001) generalized Albert's procedure to estimate parameters of the three-parameter normal ogive model and a model with multidimensional ability parameters.

Kim (2001) examined the accuracy of parameter estimates in the one-parameter logistic model using MCMC with Gibbs sampling. Four datasets were analyzed using Gibbs sampling method along with MLE methods, including conditional maximum likelihood estimation (CMLE), JMLE, and MMLE (expected a posterior method, EAP, was used to estimate θ parameters). He found that item parameter estimates from the four methods were almost identical, and θ estimates from Gibbs sampling were similar to those obtained from EAP.

Patz and Junker (1999a) applied MCMC using a Metropolis-Hastings sampling algorithm to estimate parameters for the two-parameter logistic IRT model. Later they extended this strategy to the data with multiple item formats (multiple-choice and partial credit items), missing data, and rated responses (Patz & Junker, 1999b). They demonstrated how MCMC approach is more straightforward and relatively easier to implement than MML/EM as IRT model complexity increases, since computationally MCMC does not involve exact numerical quadrature (for the E step) or pre-calculation of

derivatives (for the M step). However, the cost of this ease of implementation is that the execution time is generally slower than EM due to the fact that MCMC is trying to estimate the entire joint posterior distribution function of all the parameters (this will be introduced later) while EM only estimates one or two values for each parameter – the MML estimate and its standard error.

The examples reviewed above show some advantages of using MCMC to estimate model parameters, including its flexibility, ease of implementing for complex IRT models, and accuracy of parameter estimates. The major drawback is that it usually requires a considerable amount of computing time. Perhaps the rapid development of personal computers (PCs) will alleviate this problem. Relatively few applications (e.g., Hoijtink & Molenaar, 1997) using MCMC techniques for LC models were found in the literature.

The general form of the Bayesian approach can be written as Equation (11):

$$p(y \mid x^*) \propto p(x^* \mid y) \times p(y), \tag{11}$$

where

$p(y \mid x^*)$ is the posterior density function of y given the observed data x^*;

$p(x^* \mid y)$ is the likelihood function for x^* given y; and

$p(y)$ is the prior belief about y.

If one uses the 3-PL model as an example and lets $g(\theta_n)$, $f(a_i)$, $f(b_i)$, and $f(c_i)$ denote the prior beliefs about the ability of examinees θ_n (n=1,…, N), the item discrimination parameter a_i (i=1,…, I), the item difficulty parameter b_i (i=1,…, I), and the guessing parameter c_i (i=1,…, I), then the joint posterior density of the parameters θ, a, b, and c (i.e., the posterior to observing the item responses) is given by

$$f(\theta, a, b, c \mid x^*) \propto L(x^* \mid \theta, a, b, c)\left\{\prod_{i=1}^{I} f(a_i) f(b_i) f(c_i)\right\}\prod_{n=1}^{N} g(\theta_n). \quad (12)$$

From Equation (11), one can see that the right side of the equation is proportional to the posterior. To make the posterior a proper distribution, one must obtain the normalizing constant *C*. If *y* is a discrete variable, then

$$p(y_k \mid x^*) = p(x^* \mid y_k) p(y_k) / C, \quad (13)$$

where $C = \sum_j p(x^* \mid y_j) p(y_j)$, (14)

and the summation runs over all possible values of *y*. If *y* is a continuous variable, then

$$p(y \mid x^*) = p(x^* \mid y) p(y) / C, \quad (15)$$

where $C = \int_y p(x^* \mid y) p(y) dy$. (16)

Evaluating *C* can be exceedingly difficult to evaluate with multiple variables, as the case for 3-PL model (Equation 12). However, one can use the sampling-based approximation methods (e.g., MCMC) to resolve this without having to evaluate the normalizing constant.

The key idea of Markov chain simulation is to create a chain whose stationary distribution is a specified posterior distribution and run the simulation long enough (i.e., repeating the sampling process by starting with a possible value for each variable then drawing a sample from the updated distribution and continuing to do so) that the distribution of the current draws conditioning on the previous draws is a sufficiently close approximation to the stationary distribution. At this point, approximate distributions and summary statistics for each variable can be obtained based on these many draws.

There are two widely used Markov chain simulation methods – the Gibbs sampling and the Metropolis algorithm or the Metropolis-Hastings approximation method (see Gelman et al., 1995, pp. 320-344). We start with the Gibbs sampling, since to date most statistical applications of MCMC in psychometrics have used it.

Let X_{ij} be a response for person i to item j, θ_i be a parameter(s) for person i, δ_j be a parameter(s) for item j, ξ be a parameter(s) for distribution of θs and τ be a parameter(s) for distribution of δs. A full Bayesian measurement model can be presented by:

$$p(X,\theta,\delta,\xi,\tau) = \prod_i \prod_j p(X_{ij} \mid \theta_i,\delta_j) p(\theta_i \mid \xi) p(\delta \mid \tau) p(\xi) p(\tau). \quad (17)$$

By Bayes Theorem,

$$p(\theta,\delta,\xi,\tau \mid x^*) \propto \prod_i \prod_j p(X_{ij}^* \mid \theta_i,\delta_j) p(\theta_i \mid \xi) p(\delta \mid \tau) p(\xi) p(\tau). \quad (18)$$

Then the Gibbs sampling proceeds as follows:

- Draw values from "full conditional" distributions as shown below
- Start with a possible value for each variable in cycle 0
- In cycle $t+1$,

  For each person i, draw θ_i^{t+1} from $p(\theta_i \mid \theta_{<i}^{t+1}, \theta_{>i}^{t}, \delta^t, \xi^t, \tau^t, x^*)$.

  For each item j, draw δ_j^{t+1} from $p(\delta_j \mid \theta^{t+1}, \delta_{<j}^{t+1}, \delta_{>j}^{t}, \xi^t, \tau^t, x^*)$.

  Draw ξ^{t+1} from $p(\xi \mid \theta^{t+1}, \delta^{t+1}, \tau^t, x^*)$.

  Draw τ^{t+1} from $p(\tau \mid \theta^{t+1}, \delta^{t+1}, \xi^{t+1}, x^*)$.

The Metropolis (M) or Metropolis-Hastings (MH) algorithm can be used within Gibbs iterations when the "full conditional" distributions are not in a familiar form or cannot be sampled from directly. The basic idea of the M and MH approximation methods is that one can draw from a "proposal distribution" that one can compute and sample from.

By setting up a criterion based on the density of the proposal distribution and the target distribution at the drawn point, draws from the proposal distribution are either accepted or rejected. If they are rejected, the value of this variable in the next cycle of the Gibbs sampler remains the same. The most popular choice of the proposal distribution is the normal distribution with mean at the variable's previous value and a specified (or estimated) standard deviation. In general, as long as the distribution is defined over the appropriate range, virtually any proposal distributions will work.

The M and MH algorithms differ as to whether the proposal distribution is symmetric (for M) or not (for MH), and as to the corresponding accepting rule. Given that θ is a variable in the posterior one is interested in, θ^t is its value in cycle t of a Gibbs sampler, $p(z)$ is the full conditional for z, which includes data and most recent draws for all other variables, $q(.|\theta^t)$ is the proposal distribution with mean θ^t and a specified standard deviation (e.g., 1) and y is a draw from the proposal distribution, then the proposal distribution is symmetric if $q(y \mid \theta^t) = q(\theta^t \mid y)$. Accept y as θ^{t+1} with probability min $(r,1)$ where

$$r = \frac{p(y)}{p(\theta^t)}. \tag{19}$$

On the other hand, for MH the proposal distribution does not need to be symmetric, i.e., $q(y \mid \theta^t) \neq q(\theta^t \mid y)$, and

$$r = \frac{p(y)/q(y \mid \theta^t)}{p(\theta^t)/q(\theta^t \mid y)}. \tag{20}$$

One can easily see that if the proposal distribution is symmetric Equation (20) reduces to Equation (19); therefore one could consider the MH as an extension or generalization of M. Allowing the asymmetric proposal distribution in MH can be useful in increasing the speed

of the random walk solution afforded by a MH-within-Gibbs solution. In other words, the convergence to the stationary distribution might be faster if using MH compared to M, given the proper proposal distribution.

There are some important properties for MCMC. First of all, MCMC exhibits the known Markov property of "no memory", meaning that draws in cycle $t+1$ only depend directly on values in cycle t, not on previous cycles. Second, an indirect dependence on previous values introduces autocorrelations across cycles. That is, although the sequence of draws of a given parameters does approximate the posterior of that parameter, the values are not independent draws from the distribution. Smaller autocorrelation coefficients are preferred; the value depends on the parameterization of the model, and the amount of information in the data for a given parameter. Third, under regularity conditions (e.g., sampling can "cover the space", or can choose any point in each parameter's range), dependence on starting values is "forgotten" after a sufficiently long run. Therefore, the "burn in" cycles – the first few hundreds or thousands of draws that are to be discarded because the sampled values in those cycles are dependent upon the starting values – will not be included in calculating the summary statistics for the variable one is interested in. One can run multiple chains with over-dispersed starting points to examine if they look like they are sampled from the same stationary distribution (see Gelman et al., 1995, on convergence diagnostics).

Several computer programs have been developed for specific purposes to carry out the model parameter estimations using MCMC, including a computer program written in *MATLAB* (The Mathworks, Inc., 1996) by Albert (1992), a *FORTRAN* program written by Baker (1998), and a specialized code in *S-PLUS* (MathSoft, Inc., 1995) by Patz and Junker

(1997). Each of these programs is specific to the model the researchers were studying. For the current study, the *WinBUGS* computer program is used for estimating model parameters, assessing model convergence, and comparing the nested models under the AR model (these are discussed in the next chapter). It is an interactive Windows version of the *BUGS* program (**B**ayesian inference analysis **U**sing **G**ibbs **S**ampling, Spiegelhalter, Thomas, & Gilks, 1997) for Bayesian analysis of complex statistical models using MCMC techniques (in particular Gibbs sampling, with Metropolis steps within Gibbs for full conditionals with hard-to-calculate forms). It is a widely used freeware program and has been used to a wide range of complex problems (see those examples listed in the program manual) due to its flexibility. This is particularly useful in the current study because no other computer programs are readily available for estimating parameters for the AR model. Therefore, because of the program's flexibility along with those advantages and nice features of using MCMC over other estimation procedures as discusses earlier, *BUGS* where MCMC techniques are implemented would be the best choice to estimate parameters for the AR model in the current study.

Analysis of a Mechanics Test in Physics from the Perspective of Evidence-Centered Design

The domain of interest – the SM in CAF. As described in the earlier sections, the goal of this dissertation research is to use an appropriate psychometric model to examine data that bear on how college students learn physics concepts, in particular here Newtonian mechanics. In the past, the evaluation of a given student's performance for physics tasks was mainly based on the overall score (or the number of items he/she answered correctly), and the statistical inference one can make is whether or not the student has mastered the

content. However, this does not tell much about how students may be thinking when they respond to questions presented to them. Do they have a common model (correct or incorrect one) in their reasoning to answer questions? Does each individual not only have the Newtonian thinking but also other alternatives, and is the state of reasoning evoked dependent on the features of test items? Research in the cognitive science of learning has shown that prior knowledge or experience plays an important role when an individual is learning new things or concepts. In particular, a student's responses on the test are affected by the context each item embodies if he/she has not yet mastered the concepts the test is designed to measure. Using the CAF in the "evidence-centered" assessment design and the knowledge from previous studies in the psychology of science learning, the purpose of this research is to examine how new models and methods consistent with the psychometric tradition can be used to reveal naïve physics learners' conceptions and misconceptions about Newtonian mechanics. Such approaches – the Andersen/Rasch model in particular – are solidly grounded in statistical theory and situated in psychometric notion of modeling "noisy" responses in terms of more fundamental characteristics of students and tasks. In this case, however, those characteristics are driven by the psychology and the particular substance of physics learning: specifically, what are the propensities of a student to respond under the various approaches, and what are tasks' propensities to, by virtue of their features, provoke responses of the different approaches?

Design of the assessment – the TM in CAF. To further investigate how students are learning physics concepts – to have indications of students' understanding of them and thus provide feedback to improve instruction – the features of tasks that are needed to evoke evidence about the student variable(s) need to be taken into consideration in the task

design. Using a physics test as an example, this means that if the goal of the test is to examine students' mastery level on Newton's third law, the test should consist of questions in which the understanding of Newton's third law is required to answer items correctly. If, on the other hand, one would like to explore more deeply how students are responding to those questions, each item may only embody one contextual feature (e.g., the mass or velocity of the object as a trigger factor) in order to study each factor/feature individually. Items with multiple contextual features can have the confounding effects and prevent us from analyzing and interpreting students' misconceptions (i.e., how did they arrive at the wrong answers?).

Below we present a set of choices listed on the FMCE used to examine students' understanding about Newton's third law given different scenarios to demonstrate how the situation is set up, and the response alternatives are created, in order to provoke various misconceptions. Those questions involve collisions between a car and a truck but are mixed with different physical features – mass and velocity. Based on Newton's third law, the magnitude of the forces between the car and the truck when they collide would be the same regardless of weight and speed. However, students with incorrect physical models would believe that either mass or velocity or both can result in different magnitudes of forces between the car and the truck. Therefore, five student models – one null model, one correct model, and three incorrect models – possibly exist among students (Bao, 1999):

Model 0: Null Model (i.e., nonscientific reasoning).

Model 1: Both car and truck exert the same amount of force on the other regardless of either mass or velocity. (Correct)

Model 2: The car and the truck can exert unequal amount of force on the other, and the one exerting the larger force depends on the velocity only (i.e., regardless of the subject's mass). (Incorrect)

Model 3: The car and the truck can exert unequal amount of force on the other, and the one exerting the larger force depends on the mass only (i.e., regardless of the subject's velocity). (Incorrect)

Model 4: The car and the truck can exert unequal amount of force on the other, and the one exerting the larger force depends on both the velocity and the mass. (Incorrect)

In order to sort out which of these models a student might be using under various conditions, the FCME is tacitly using a task model with a common stimulus situation – truck and car colliding head-on – and introducing variation with respect to the following task-model variables and possible values of them:

- Mass of vehicles: Same; truck heavier; and car heavier.
- Velocity: Both moving at the same velocity; both moving, and car moving faster; both moving, and truck moving faster; truck moving but car still; car moving but truck still.

(Additional task model variables that could be introduced to explore other misconceptions associated with Newton's third law are whether colliding objects are animate, are capable of intentionality, and have been moving prior to the scenario.)

Further, the students' work products are choices of proffered explanations that are provided in multiple choice format, designed to reveal thinking along the lines of the student models listed above. Specifically, seven choices on the FMCE are given to students to let them choose the answer that best describes the size (magnitude) of the forces between the car and the truck under several conditions (e.g., on question 30, students are asked to choose the best answer given that the truck is much heavier than the car and they are both moving at the same speed when they collide):

A. The truck exerts a larger force on the car than the car exerts on the truck.
B. The car exerts a larger 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 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.
G. None of the answers above describes the situation correctly.

Comparing these choices with those models students possibly use for problem-solving, it is easy to see that choices C, F, and G correspond to Model 0 (the null model – in other cases, many of the responses of which correspond roughly to an Aristotelian conception of force and motion), choice E is based on Model 1 (the correct model), and the other choices are derived from either Model 2, Model 3, or Model 4 (incorrect student models, containing some notion of force and motion, but often not quite correct in ways that can be thought as using "impetus theory" rather than Newtonian mechanics), depending on how the situation

is set up. If using question 30 – the truck is much heavier than the car but they are moving at the same speed when they interact – as an example, it should be noted that, however, students using the Model 2 – the force the subject exerts to the other depends on the velocity only – would choose the correct answer (the choice E) as well. As a result, student responses on this group of FMCE questions cannot be coded with item-based modeling as suggested by Bao (that is the same reason we exclude those questions on the FMCE in our current study – the AR model is appropriate to be used only for the data coded at the item level). This is somewhat related to the data analysis discussed in the following subsection. Our purpose here is to demonstrate how the problem situation is set up and how those choices are created to provoke students' conceptions/misconceptions. On the other hand, one can see that how we analyze the data is related to how the task is designed to reveal students' cognitive process in problem-solving.

In order to set the stage for the AR analyses we employ and to summarize previous analyses, the next section reviews Bao and Redish's analyses. In terms of the ECD framework, Bao and Redish's analyses correspond to the statistical component (or the measurement component) of the EM in the CAF.

Summary of Bao/Redish analyses – statistical analysis in the EM of the CAF. Two methods were developed by Bao and Redish (2001 & 2004) in their studies on students' physics learning: *Concentration Analysis* and *Model Analysis*. They first developed *Concentration Analysis* to measure how students' responses on multiple-choice questions are distributed. This information can be used to explore if the students have common correct/incorrect models or if the question is effective in detecting models of students' reasoning for problem-solving. Suppose a multiple-choice single-response

question with 5 choices is given to 100 students. Examples of distributions of responses could be: (1) the responses are evenly distributed among all the choices (i.e., 20 students for each choice), implying random guessing; (2) there is a higher concentration on some choices than on others, which is a more typical distribution that may occur in our classes; and (3) only one choice is selected by all students, giving a 100% concentration. Based on this example, the concentration factor (C) is defined "as a function of student response that takes a value in [0,1]." A larger value of C indicates more concentrated responses with 1 being a perfectly correlated response (the 3rd type of the distribution above) and 0 a random response (the 1st type of the distribution above). The value of C would be between 0 and 1 for the 2nd type of distributions. This concentration factor can be calculated using Equation (21),

$$C = \frac{\sqrt{m}}{\sqrt{m}-1} \times \left(\frac{\sqrt{\sum_{i=1}^{m} n_i^2}}{N} - \frac{1}{\sqrt{m}} \right), \tag{21}$$

where m is the number of choices, n_i is the number of students select the choice i, and N is the total number of students.

Several methods of using the concentration factor were introduced, and results from the FCI (Hestenes et al., 1992) were used to demonstrate them. For example, one can cross tabulate the concentration factor with scores (or the percentage of items students answer correctly) if both of them are recoded into categories based on criteria set by researchers. This can be used to show if the questions trigger some common "misconceptions." Assume that students' scores and the concentration factors are categorized into three levels (*L*ow, *M*edian, and *H*igh). Then the following patterns are

meaningful and provide some information about the student data: *HH* (one correct model), *LH* (one dominant incorrect model), *LM* (two possible incorrect models), *MM* (two popular models, one is correct and another is incorrect), and *LL* (random guessing). Other methods involve constructing plots, and they can provide additional information about how students' responses shift from pre-instruction to post-instruction (if data from both pre- and post- instruction is available) or information about incorrect answers.

To look for the detail of those possible situations of student models for reasoning in physics learning, Bao and Redish (2004) developed the second method, *Model Analysis*, to extract the probability states of students' use of different models. The analysis is mainly based on the cognitive science of learning (e.g., context dependence, as discussed earlier) and the knowledge from qualitative research (i.e., interview students to find out possible contextual features in learning the Newtonian mechanics). It involves two important concepts which have been mentioned but not well-defined in the preceding sections: *common models* and *student model states*. The *common models* are those models commonly used by students. They often consist of one correct expert model and a few incorrect or partially correct student models. When students are presented with a set of questions for a new physics concept or situation, they may respond using a previously well-formed model or create a new model based on their past experience or knowledge (e.g., mapping of a reasoning primitive). As an example, with the concept of the force-motion relation in the Newtonian mechanics, there are three common models students may use as indicated on the Bao and Redish's paper:

Model 1: An object can move with or without a net force in the direction of motion. (an expert model using Newtonian way of thinking)

Model 2: There is always a force in the direction of motion. (an incorrect student model using "impetus theory" belief)

Model 3: Null model. (an unsystematic, inconsistent, or Aristotelian approach)

Note that in practice, there can be more than one more specific model corresponding to each of the model categories described above. In this dissertation, response options that corresponded to the same model category will be collapsed into a single category (more details are described in the following chapter). For example, suppose a task provide six multiple choice options, and two were consistent with a correct student model, three with an incorrect student model, and one with a null model. For the purposes of Bao and Redish's analysis and for the AR analysis of this dissertation, the first two would be combined into one common Newtonian category and the next three would be combined into a single category thought of as using "impetus theory" approach .

Students may consistently use one of the common models (correct or incorrect one) to answer all questions or they may inconsistently use different common models depending on what model is triggered by the given item. These different situations of using models when naïve students are presented test questions related to a new concept are described as *student model states*. The first case above (i.e., consistently use the same model) is called a pure model state while the second case (i.e., inconsistently use different models) is referred to a mixed model state. A mixed model state is common when students are in the transition of mastering a new concept.

The idea of *Model Analysis* is that if one can design a set of questions to probe a particular concept, the probability for a given student to activate the different common models in response to these questions can be measured appropriately. That is, a student's model states can be represented by a set of probabilities. Suppose a population of students is given *j* multiple-choice single-response questions on a single concept for which this population uses *m* common models. The k^{th} student's probability distribution measured with the *j* questions can be represented by a vector space, $\vec{V}_k$:

$$\vec{V}_k = \begin{pmatrix} v_1^k \\ v_2^k \\ \vdots \\ v_m^k \end{pmatrix} = \frac{1}{j} \begin{pmatrix} n_1^k \\ n_2^k \\ \vdots \\ n_m^k \end{pmatrix}, \tag{22}$$

where v_ω^k is the probability of the k^{th} student being in the ω^{th} model state and equals to $\frac{n_\omega^k}{j}$; n_ω^k represents the number of questions in which the k^{th} student used the ω^{th} common model; and

$$\sum_{\omega=1}^{m} \mathrm{n}_\omega^{\mathrm{k}} = j\,. \tag{23}$$

The Equation (22) can be represented in a square root form to have a unit vector, $\vec{u}_k$, in the model space:

$$\vec{u}_k = \begin{pmatrix} \sqrt{v_1^k} \\ \sqrt{v_2^k} \\ \vdots \\ \sqrt{v_m^k} \end{pmatrix} = \frac{1}{\sqrt{j}} \begin{pmatrix} \sqrt{n_1^k} \\ \sqrt{n_2^k} \\ \vdots \\ \sqrt{n_m^k} \end{pmatrix}. \tag{24}$$

A procedure called *model estimation* is then introduced to extract information about students' use of models in terms of eigenvalues and eigenvectors derived from the *class model density matrix* based on the *single student model density matrix*. Suppose *m* equals 3 (i.e., involves 3 common models). The *single student model density matrix* is defined as in Equation (25), a product of $\vec{u}_k$ and $(\vec{u}_k)^T$:

$$\Phi_k = \frac{1}{j}\begin{bmatrix} n_1^k & \sqrt{n_1^k n_2^k} & \sqrt{n_1^k n_3^k} \\ \sqrt{n_2^k n_1^k} & n_2^k & \sqrt{n_2^k n_3^k} \\ \sqrt{n_3^k n_1^k} & \sqrt{n_3^k n_2^k} & n_3^k \end{bmatrix}. \tag{25}$$

Then the *class model density matrix* is defined as the average of the individual students' model density matrices as in Equation (26):

$$\Phi = \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} \\ \delta_{21} & \delta_{22} & \delta_{23} \\ \delta_{31} & \delta_{32} & \delta_{33} \end{bmatrix} = \frac{1}{N}\sum_{k=1}^{N}\Phi_k , \tag{26}$$

where *N* is the number of students.

It can be shown that the diagonal elements of Φ reflect the percentages of students' responses for each common model used while the off-diagonal elements reflect the consistency of the model used by students. Suppose δ_{11} equals a non-zero number and other elements in Φ equal zero. It indicates that students consistently use one model (correct or incorrect one). If δ_{11}, δ_{22}, and δ_{33} (i.e., the diagonal elements) equal some non-zero numbers and the off-diagonal elements are all zeros, implying that three common models are consistently used by students. On the other hand, if all elements in Φ are not zeros, implying that students use three models inconsistently, and large off-diagonal elements indicate low consistency for individual students in their model use.

Furthermore, an eigenvalue decomposition method can be used to extract the class model vectors (the eigenvectors of Φ) and the eigenvalues. As discussed in Bao and Redish's paper, these eigenvectors representing the class model states reflect "the salient features of all the individual student model vectors." In the case where students have two dominant models (a correct one and a common misconception), one can construct a two-dimensional graph or "model plot" using the eigenvectors corresponding to those two model states to represent the student usage of the two models. In terms of the eigenvalues, it can be shown that they are affected by two factors: the similarity of the individual students' model vectors and the number of students with similar model state vectors. Therefore, a large eigenvalue obtained from Φ implies majority of students in the class tend to have similar single student model vectors. That is, most students use the same common model. Compared to *Concentration Analysis*, this quantitative information tells us clearly about which common model the majority of students use – it can be Model 1, Model 2, or Model 3 if using the example above. On the other hand, several small eigenvalues indicate that students apply their models differently from each other. Data from FCI (5 questions that activate models associated with the force-motion concept and student responses were coded according to those three common models as seen in the example) was used to demonstrate the *Model Analysis*. Later, Bao, Hogg, and Zollman (2002) showed the effectiveness of *Concentration Analysis* and *Model Analysis* on the self-designed questions following what cognitive psychology suggests about students' science learning to evaluate college students' understanding about Newton's third law.

Psychometric analyses. Methods developed by Bao and Redish as described above provide a better way than the traditional, total score, evaluation method in revealing how students learn new physics concepts in Newtonian mechanics. It helps to design a more valid instrument and to improve instruction in physics. However, there are some limitations in their methods.

First, it is not clear whether these methods can be applied to other fields of learning (e.g., mathematics). The cognitive learning process for mathematics may be different from learning physics since they are two different subjects (i.e., their substances are not the same). More generalized analyses would be preferred in educational testing. For example, psychometric models which are statistical models are substance-independent, yet when appropriately constructed and applied, reflect the key patterns in the substantive problem at hand. Thus analyses based on them can be used in various subjects of learning while remaining true to the learning theory of the domain of interest.

Second, the Boa-Redish analyses are not connected with the well-developed psychometric machinery, where much has been learned over the past century about issues such as estimation, model criticism, and modeling approaches. Some measurement models (e.g., IRT models) have been widely used in educational testing. Although they may not be sufficient for Bao and Redish's interests in knowing how students learn new physics concepts, the little-known AR model as described above does exist, and it is consonant with the conception of student model states in the Bao-Redish approach. Their analysis is based on the belief that naïve students use different models, with probabilities that depend in part on the features of the test items. This could be described as a mixture-within-persons model as the AR model. Posterior probability vectors associated

with the different model states in the AR model would reveal the same kinds of patterns the Boa-Redish eigenvalues summarize.

Third, their analyses are data dependent, in the same sense as those of the classical test theory. That is, all of statistics in their analyses will vary if different sets of questions are given to students. For example, v_{ω}^{k} in Equation (22) will be changed if using a different set of questions. That is, student model states parameterized in this way are not unique for a given concept as they indicated in their paper, "the student model state represents an interaction between the student and particular instrument chosen." (Bao & Redish, 2004, p. 11). This limits the test use, and it does not allow comparing students' performance if they have taken quite different subsets of test items. In modern test theory, like the AR and other IRT models, once the assumptions of the model are satisfied (and they can be examined using statistical procedures), the item (or person) parameter estimates are independent of the particular sample of students (or test items) (Hambleton & Swaminathan, 1985). Thus different subsets of the same item pool would yield different concentration analysis estimates but statistically equivalent AR item parameter estimates, even while both models were faithful to the same mixture-within-persons response patterns. (This aspect of the analysis will be addressed directly in Chapter IV, in a comparison of results of a four-item test and an eight-item test that includes the original four.)

Next, neither of their methods provides estimates and accompanying measures of accuracy at the level for individual students, as well as standard error of estimate at the item level. These are very important and essential in educational measurement (no matter whether the use is for classroom assessments or high-stake tests). *Concentration Analysis*

is mainly used to examine the effectiveness of the test items (but without providing accompanying standard errors of estimation), and it is not intended to measure students' mastery or propensity levels. *Model Analysis* is based on the class model density matrix and is used to evaluate two types of consistency – the consistency of individual students using different models by examining the structure of the class model states (mixed or pure) and the consistency among different students reflected by the eigenvalues. Analyses based on psychometric models provide more useful and important information both at the individual and item levels than those two analyses, yet also fulfill Bao and Redish's intentions to know how individual students respond to physics questions. Besides, in most cases, psychometric analyses simultaneously produce the item and person parameter estimates, whereas two separate methods are needed in Bao and Redish's studies to conduct the item or person level analysis. Furthermore, psychometric analyses support associated standard errors of estimation that are not available in Bao and Redish's analyses. There is no clear way to examine the precision of estimated statistics in their analyses. Thus the AR model is a good example of the psychometric approach and its attendant advantages to a test built around contemporary, rather than traditional, views of learning.

Finally, there are no procedures developed for statistical model fitting and model comparisons in Bao and Redish's analyses. They provide ingenious descriptive tools to study students' physics learning in great detail; however, lack of statistics for model fitting and model comparisons prevents the researcher from assessing the effectiveness of their analyses and comparing with other alternatives. Bao and Redish showed the effectiveness of their methods using data from FCI, but it was based on some qualitative information and

was mainly targeting physics. In terms of a statistical point of view, it would be better to have a quantitative index for describing how well the model fits the data. They also compared *Model Analysis* with *Factor Analysis* using two special cases and showed that the former is more valid than the latter in detecting the class model states. Again, a statistical index for model comparisons would be recommended rather than just showing the model differences by using two extreme cases. In Chapter III, we introduce a Bayesian model-fit index that was used in the current study to assess the fit of the model and to compare various models under the AR model.

The Bao-Redish model analyses provide an innovative approach to examining data from a rich, substantively- and cognitively-grounded set of test items. Appreciating the patterns they seek to model yet recognizing the limitations of the methods, psychometric analyses based on the AR model are conducted in the current study. The general model description and its inference on the student variable have been discussed in the earlier sections. (Although the 3-PL model and LC models are not applied, they were contrasted in a previous subsection with the AR model in the way they model students' responses, based on the psychological perspective they represent and the inference they can make about students' learning.)

Summary

As briefly discussed in the very beginning and developed throughout in the paper, this dissertation research is meant to provide an example of analyses integrating ideas from several areas of current research. First, it is based on the ECD framework developed by Mislevy, et al (2003). Within this framework, this dissertation focuses on the measurement component of the evidence model in the CAF. Second, as the primary goal, it is desired to

compare and explore the utility of analyses that draw upon the armamentarium of psychometrics to make inferences about the student variable(s) using data from physics as an example. Bao and Redish (2001 & 2004) have developed both the *Concentration Analysis* and *Model Analysis* to study college students' learning in physics (especially in Newtonian mechanics). Their analyses are shown to be effective but with the limitations discussed above.

Next, this line of research also integrates findings from the psychology of science learning with psychometric methods. One of major findings that Bao and Redish draw upon is that naïve students' responses on questions are affected by their pre-existing knowledge or experience (i.e., context dependence). Therefore, task questions designed to measure students' understanding of physics concepts are suggested to embody the contextual features associated with each item targeting on the specific concept. Bao and Redish (2001 & 2004) have showed that five questions on FCI are effective in evoking features related to the force-motion concept in Newtonian mechanics through their analyses.

Finally, parameter estimation for the AR model can be carried out by using MCMC techniques. The computer program, *WinBUGS*, is available for doing this kind of task, and will illustrate the use of MCMC estimation procedures with an innovative model and application through this study.

Chapter III

Methodology

Data

Four data sets were used: (1) the FCI data with 5 items (questions 5, 9, 18, 22, and 28) and 198 subjects; (2) the Force-Motion data , containing 403 examinee responses to 4 items (questions 2, 5, 11, and 12) from the FMCE; (3) the Force-Motion data, obtained from the same subjects as for the second data set but with 4 additional items (i.e., questions 2, 5, 11, and 12 plus items 8, 9, 10, and 13); and (4) the Acceleration data, also obtained from the same subjects as for the second data set, containing students' responses on questions 22 through 26. All of these questions are listed in Appendix A.

As indicated earlier, the FCI and FMCE are the two most commonly used instruments in physics to measure students' understanding on concepts in Newtonian mechanics. Items on the first two data sets have been used by Bao and Redish in their studies (Bao, 1999; Bao & Redish, 2001, 2004). Therefore, the findings from the current study can be used to contrast with what they found in general terms (since the datasets are different). For the following two reasons, we also conducted analyses using the third data set. First, we can compare the parameter estimates for those questions used both in the second and third data sets – since they are the same items, it is expected that the parameter estimates would be similar, with some variations due to sampling error. Second, we can examine whether more items would yield more stable estimates in the case of the third data set. (In the context of IRT, more items and/or more examinees are preferred in order to obtain more stable and accurate estimates). We employed the fourth data set because

acceleration is also an interesting topic in physics. In addition, it gives us an opportunity to explore whether the mapping schema created by Bao and Redish and mainly used on the force-motion concept can be applied to other concepts (if applicable) as well.

The FCI data were collected from the algebra-based physics course (PHYS 121) taught in the Fall of 2001 at the University of Maryland. Most of the students were from the College of Biology (about 70-80% typically) and most were juniors or seniors. They were given the FCI in the first and last weeks of the class. Therefore, the data contain students' responses on both pre- and post-tests. This allows us to examine whether students make some progress in understanding the force-motion relation after one semester, as well as to explore the homogeneity of the item parameter estimates. This can be done through model comparisons that are described later in this chapter.

The FMCE test was given to a population similar to that for the FCI data. They were collected in the first and last weeks of the class during the Fall of 2000 and 2002. Since the same instrument was given to similar populations, the data from those two years are combined in this study. In addition, as for the FCI data, the FMCE data also contain students' responses from pre- and post-tests that are distinguished in the analysis. As mentioned earlier two concept groups measured in FMCE are considered in the current study: the Force-Motion concept (for the second and third data sets) and the Acceleration concept (for the fourth data set).

Analyses

The responses from all four data sets (both pretest and posttest) were originally coded based on five choices. Bao (1999) and Bao and Redish (2004) recoded the students' responses for the first two data sets into three categories based on three student models, namely Newtonian (Model 1), "impetus theory" (Model 2), and a "null" or Aristotelian (Model 3) conception. Therefore, in this study, responses from the first two data sets were recoded into three response categories based on this coding scheme. We also recoded students' responses on both the third and fourth data sets following Bao and Redish's mapping strategy. These codings can be found in Appendix B. Then the recoded responses would be adequate to be analyzed for the AR model. It should be noted that for the first data set there is no response category 3 (i.e., no responses would be recoded under category 3) for question 22. However, to be consistent with other items, we use the model as if the response category 3 exists for this item when estimating model parameters and was not used by any student (and we anticipate a parameter estimate that indicates a response in category 3 is very unlikely!).

After the data were recoded, standard descriptive item analyses were conducted for each data set, including a frequency distribution table for each question based on the original five response categories and the three response categories collapsed in terms of physics conceptions. In addition, the Pearson correlation among items for each data set that are based on the three response categories were performed, followed by the polyserial correlation analyses (between the item and students' overall scores) also based on the three response categories. (Even though the categories 1-3 are, strictly speaking, nominal, they can be ordered from 1 representing a higher level of understanding than 2, which is again

higher than 3.) The analyses first were conducted for pre- and post-test separately then they were analyzed using the combined data set (i.e., combine the pre- and post-test for each data set). These analyses were used to help to identify some possible mistakes resulting from data entry and to provide some basic information about items.

Each data set then was analyzed under the AR model, using the computer program *WinBUGS*. Missing data in *WinBUGS* are treated as "missing at random" (MAR) – that is, the distribution of the missing-data mechanism does not depend on the missing values, rather it is permitted to depend on other observed values through the proposed model (Gelman, et al, 1995). As mentioned beforehand, in order to examine whether the item parameters and/or the person parameters are homogeneous with respect to testing occasion, three models under the AR model were compared. This modeling strategy is analogous to a common strategy in latent class modeling (e.g., Dayton, 1998, p. 78). If the item parameter estimates are homogeneous, this indicates that each item has a similar tendency to evoke the specific conception/misconception regardless of time points; likewise, the homogeneity of the person parameter estimates implies that student populations' propensity distributions, reflecting their understanding about physics concepts, do not change from the pre-test to the post-test. These three models are:

First, the homogeneous model: One *BUGS* run, with the same conditional probabilities for items and the same examinee population distributions for θs over all subjects and time points.

Second, the partially homogeneous model: One *BUGS* run, with the same conditional probabilities for all subjects and time points but different population distributions for pre-test response data and post-test response data. In *BUGS*, this is

accomplished by including a data variable time (1 for pre-test and 2 for post-test, for instance) for each data vector, and having distinct examinee distributions for the two sets. The θs for the pre- and post-test would come from normal distributions but with different means, and they can be estimated empirically.

Third, the heterogeneous model: One *BUGS* run again, with same distinct pre- and post-test distributions for subject as in the partially homogeneous model but now with two sets of item parameters for each item (i.e., one for the pre-test and another for the post-test). The time-point variable associated with each data vector determines which set of item parameters is used with that data vector.

The *BUGS* code for each condition is listed in Appendix C. Since those codes are identical for each data set (they only differ mainly on the data vectors), only the ones for the first data set are presented.

Each *BUGS* run consists of the following steps: model specification (select *specification* under **model** menu), sample monitor (select *samples* under **inference** menu), and sample update (select *update* under **model** menu). Through model specification, one can examine whether the program code for the full Bayesian model in question is syntactically correct, load data, compile the model with number of MCMC chains the user specifies, and load initial values for parameters that need to be estimated or let the program generate the initial values. The sample monitor tool is to monitor the nodes (the variable of interest, i.e., the parameters being estimated) by specifying *begin* and *end* – numerical values used to select a subset of the stored sample for analysis (by default, the value for *end* is 1,000,000), and *thin* – used to select every *k*th iteration of each chain to contribute to the statistics being calculated, where *k* is the value for *thin* that a user may specify. The sample

update tool allows one to specify the number of iterations desired for each run, and it is used to continue the *BUGS* run until it reaches convergence.

Checking convergence is necessary, and can be carried out by several ways (they are options under the sample monitor tool). Note that convergence in MCMC estimation is not convergence to point estimates, but rather to draws from a stationery distribution; in particular, the posterior distribution for each parameter in the model. Thus different values are obtained in each cycle, and the issue is whether they can be considered to be draws from a stable underlying distribution (Gelman, 1996).

Most convergence diagnostics and in particular those included in *WinBUGS* involve running multiple chains, with each chain starting from a distinct set of initial values for each parameter being estimated (usually, the item parameters). The first convergence-monitoring approach is to examine trace plots of the sample values versus iteration for each chain and see if they are mixing well (see Figure 4 for a reasonable convergence from two chains). Second, one can look at history plots, a complete trace for the variable being monitored (see Figure 5 for an acceptable convergence). Finally, one can examine the Gelman-Rubin convergence statistics (Brooks and Gelman, 1998). The Gelman-Rubin statistic (or is called "R") would general be expected to greater than 1 if the initial values are adequately over-dispersed between two chains and would be approximately equal to 1 if reaching convergence. In addition, one should also pay attention on examining whether the convergence is stable (see Figure 6 as an example). The green line on the figure represents the width of the central 80% interval of pooled runs in bins of length 50. The blue line, on the other hand, represents the average width of the 80% intervals within the individual runs, again in the bins of length 50. Their ratio (=

pooled / within), the R-statistic, is represented by the red line. In this study, each *BUGS* run consists of two chains to check for the convergence of the chains to a common stationary distribution.

Figure 4.
A Trace Plot for a Reasonable Convergence from Two Chains

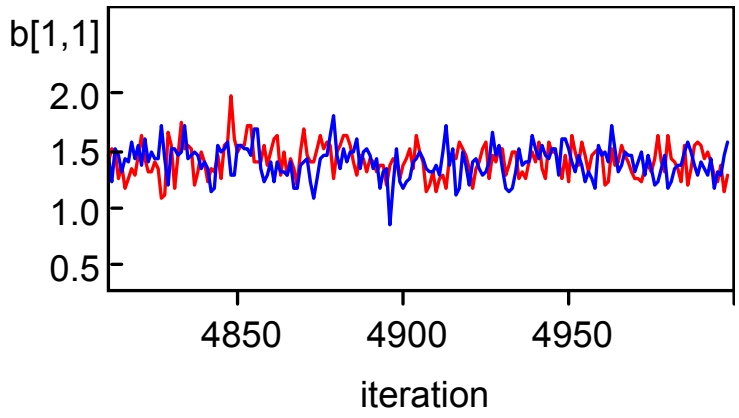


Figure 5.
A History Plot for an Acceptable Convergence from Two Chains

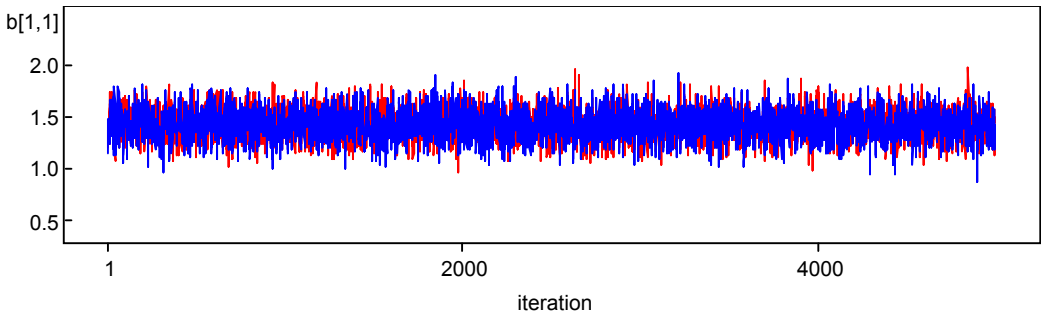
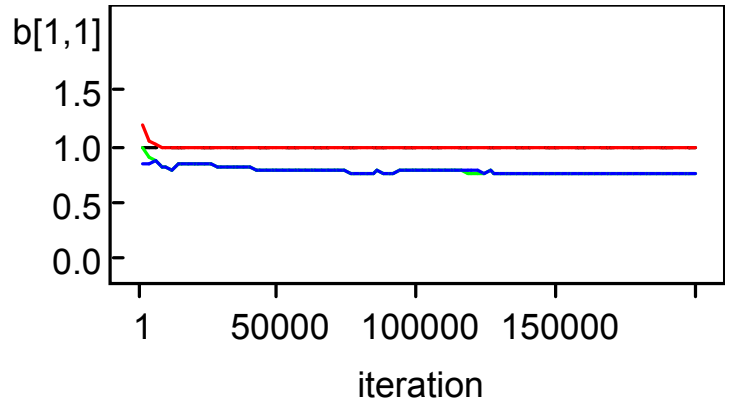


Figure 6.
The bgr Diagnosis – the Gelman-Rubin Statistic Represented by the Red Line



After convergence, a further number of iterations need to be run to obtain samples that can be used for posterior inference (i.e., to obtain the summary statistics – e.g., the mean, standard deviation, and quantiles – of the posterior distributions for the parameters being estimated). In general, more iterations (i.e., more samples from the posteriors) would produce more accurate posterior estimates. One way to assess the accuracy of the posterior estimates is to compare the Monte Carlo error (the MC error) with the sample standard deviation (SD) for each parameter of interest. SD is an estimate of the uncertainty of estimation of the parameter due to having finite data, and its size is determined by the data and the model. MC error is an estimate of the uncertainty due to having only a finite number of draws in the MCMC chains, and it can be driven to zero simply by running long enough chains. As a rule of thumb, it is suggested that the simulation should be run until the MC error for each parameter being estimated is less than about 5% of the sample SD. Since the MC error and sample SD are reported in the summary statistics table, the comparison between those two statistics can be done easily.

The same samples being used to obtain the summary statistics were also used to compute a statistic called the Deviance Information Criterion (*DIC*; Spiegelhalter, Best, Carlin, & van der Linde, 2002) and related statistics that are described below. The *DIC* is provided and calculated by *BUGS*. It is appropriate for comparing both nested and non-nested models in which the effective number of parameters being estimated is not clearly defined (due to the incorporation of prior distributions and hierarchical model structures).

This is the case here. In the current study, we would like to compare different models (i.e., the homogeneous, partially homogeneous, and heterogeneous models) under the AR model, and the parameter estimates for both the partially homogeneous and heterogeneous models involve setting up some hyperparameters (e.g., mu11 and mu12 in the *BUGS* code, the means of prior distributions on the examinee population distribution means). Furthermore, the inclusion of a prior distribution (also applied in our analyses) induces a dependence between parameters being estimated which in turns to reduce the effective dimensionality to some extents (Spiegelhalter et al. 2002). For this reason, other information criteria such as *AIC* (Akaike, 1973 & 1987) and *BIC* (Schwarz, 1978) may not be appropriate for model comparisons. The computation of both *AIC* and *BIC* depends on a measure of the effective number of parameters in the model, and they are not clearly defined in hierarchical Bayesian analyses such as ours. For these reasons, the *DIC* was used in the current study to compare different models. The definition and general idea of *DIC* is described below.

Model comparisons by DIC. The *DIC* is given by:

$$DIC = \overline{D} + pD = \hat{D} + 2 \times pD \ . \qquad (27)$$

$\overline{D}$, suggested as a Bayesian measure of fit or adequacy, is the posterior mean of the deviance where deviance is defined as -2 * log (likelihood). The likelihood is defined as $p(y \mid \theta)$, where y comprises all stochastic nodes given values (i.e. data), and θ comprises the stochastic parents of y – "stochastic parents" are the stochastic nodes upon which the distribution of y depends, when collapsing over all logical relationships. In general, the value of $\overline{D}$ decreases when fit increases. *pD*, suggested as a measure of model complexity, and is given by $pD = \overline{D} - \hat{D}$ where $\hat{D}$ is a point estimate of the deviance (i.e., -2 * log (likelihood)) obtained by substituting in the posterior means, $\overline{\theta}$, thus $\hat{D} = -2 \times \log(p(y \mid \overline{\theta})$. As we can see, *pD* is the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest. Meng and Rubin (1992) showed that such a difference is the key quantity in estimating the degrees of freedom of a test. In simple models (e.g., a simple regression model), *pD* is the effective number of parameters. This interpretation does not necessarily apply to more complex situations such as the analyses carried out here. (For example, the value of *pD* will not necessarily increase as the fitting model seemingly becomes more complex, since the effect of priors and the configuration of the data may effectively mean that either more or less is being demanded from the data in the way of inference.)

From Equation (27) one also can see that *DIC* as a type of Bayesian criterion used to compare models combines both measure of fit ($\overline{D}$) and complexity (*pD*). This approach is different from *AIC* and *BIC*. On the other hand, like *AIC* or *BIC,* the model with the smallest *DIC* is gauged to be the model that would best predict a replicate dataset of the

same structure as that currently observed. However, it could be misleading just to report the model with the lowest *DIC* if the difference is, say, less than 5, and the models imply very different inferences. In general, models with the difference within 0 to 2 deserve similar consideration, and models with *DIC* values greater by an amount of 4 to 7 have notably less support. These rules of thumb, suggested by Burnham and Anderson (2002) and commonly applied with *AIC,* appear to work reasonably for *DIC* as well (Spiegelhalter et al. 2002). As with *AIC* and *BIC*, *DIC*s are comparable only for models with exactly the same observed data, but there is no requirement for them to be nested.

Chapter IV

Results and Discussion

This chapter begins with descriptive analyses for the items, which is followed by the model-based analyses.

Descriptive Item Analyses

Tables 1 through 4 show frequency distributions based on the original response categories for each item on the pre-test, post-test, and combined test for each data set, respectively. Tables 5 through 8, on the other hand, show the frequency distribution based on the three response categories (collapsed to Newtonian, impetus theory, and null groupings) for each data set, respectively. Note that for the first data set, one examinee took the pre-test only. The student might have dropped out sometime during the semester or was absent from the class when the test was administered. His/her responses on the pre-test were included in the analyses. Therefore, there are 99 respondents in the pre-test while only 98 respondents in the post-test, and the total number of respondents is 197. For the fourth data sets, three examinees who did not answer any of questions either on the pre- or post-test were excluded in the analyses.

The Pearson correlations among items as well as between the items and the test score for each data set (based on the three response categories) in terms of pre-test, post-test, and combined data is listed in Tables 9 through 12, respectively. While the AR model does not require ordered data, as is assumed by using these descriptive statistics, it is true that in the coded data, model 1 responses are generally more desirable than model 2 responses, which are in turn more desirable than model 3 responses.) These results are

followed by the polyserial correlations between the items and the test score for each data set (also in terms of pre-test, post-test, and combined data) as seen in Tables 13 through 16.

Table 1.
Frequency Distribution Based On the Original Five Response Categories for the 1st Data Set

Response	Pre-test					Post-test				
Category	5	9	18	22	28	5	9	18	22	28
a	7 (7.1)	5 (5.1)	14 (14.1)	26 (26.3)	3 (3.0)	4 (4.1)	11 (11.2)	3 (3.1)	19 (19.4)	1 (1.0)
b	19 (19.2)	23 (23.2)	10 (10.1)	34 (34.3)	3 (3.0)	51 (52.0)	21 (21.4)	55 (56.1)	47 (48.0)	3 (3.1)
c	25 (25.3)	25 (25.3)	27 (27.3)	2 (2.0)	6 (6.1)	8 (8.2)	20 (20.4)	4 (4.1)	5 (5.1)	7 (7.1)
d	19 (19.2)	7 (7.1)	47 (47.5)	30 (30)	51 (51.5)	28 (28.6)	10 (10.2)	24 (24.5)	24 (24.5)	6 (6.1)
e	29 (29.3)	39 (39.4)	98 (99.0)	3 (3.0)	31 (31.3)	7 (7.1)	36 (36.7)	12 (12.2)	2 (2.0)	80 (81.6)
missing	0 (0.0)	0 (0.0)	1 (1.0)	4 (4.0)	5 (5.1)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.0)	1 (1.0)
Total	99	99	99	99	99	98	98	98	98	98

Table 1 (continued).
Frequency Distribution Based On the Original Five Response Categories for the 1st Data Set

Response	Combined				
Category	5	9	18	22	28
a	11 (5.6)	16 (8.1)	3 (1.5)	45 (22.8)	4 (2.0)
b	70 (35.5)	44 (22.3)	69 (35.0)	81 (41.1)	6 (3.0)
c	33 (16.8)	45 (22.8)	14 (7.1)	7 (3.6)	13 (6.6)
d	47 (23.9)	17 (8.6)	51 (25.9)	54 (27.4)	57 (28.9)
e	36 (18.3)	75 (38.1)	59 (29.9)	5 (2.5)	111 (56.3)
missing	0 (0.0)	0 (0.0)	1 (0.5)	5 (2.5)	6 (3.0)
Total	197	197	197	197	197

Note. The number in the parenthesis represents the percentage associated with that frequency count.

Table 2.
Frequency Distribution Based On the Original Response Categories for the 2nd Data Set

Response	Pre-test				Post-test			
Category	2	5	11	12	2	5	11	12
A	3 (1.5)	9 (4.4)	13 (6.3)	10 (4.9)	5 (2.5)	1 (0.5)	96 (48.7)	94 (47.7)
B	179 (87.3)	94 (45.9)	9 (4.4)	5 (2.4)	102 (51.8)	51 (25.9)	13 (6.6)	9 (4.6)
C	7 (3.4)	3 (1.5)	1 (0.5)	0 (0.0)	9 (4.6)	7 (3.6)	11 (5.6)	5 (2.5)
D	13 (6.3)	71 (34.6)	1 (0.5)	176 (85.9)	79 (40.1)	126 (64.0)	6 (3.0)	84 (42.6)
E	0 (0.0)	1 (0.5)	15 (7.3)	7 (3.4)	0 (0.0)	1 (0.5)	10 (5.1)	3 (1.5)
F	2 (1.0)	25 (12.2)	40 (19.5)	1 (0.5)	0 (0.0)	8 (4.1)	6 (3.0)	1 (0.5)
G	1 (0.5)	0 (0.0)	125 (61.0)	4 (2.0)	0 (0.0)	2 (1.0)	55 (27.9)	1 (0.5)
J	0 (0.0)	1 (0.5)	0 (0.0)	0 (0.0)	2 (1.0)	1 (0.5)	0 (0.0)	0 (0.0)
missing	0 (0.0)	1 (0.5)	1 (0.5)	2 (1.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Total	205	205	205	205	197	197	197	197

Table 2 (continued).
Frequency Distribution Based On the Original Response Categories for the 2nd Data Set

Response Category	Combined			
	2	5	11	12
A	8 (2.0)	10 (2.5)	109 (27.1)	104 (25.9)
B	281 (69.9)	145 (36.1)	22 (5.5)	14 (3.5)
C	16 (4.0)	10 (2.5)	12 (3.0)	5 (1.2)
D	92 (22.9)	197 (49.0)	7 (1.7)	260 (64.7)
E	0 (0.0)	2 (0.5)	25 (6.2)	10 (2.5)
F	2 (0.5)	33 (8.2)	46 (11.4)	2 (0.5)
G	1 (0.2)	2 (0.5)	180 (44.8)	5 (1.2)
J	2 (0.5)	2 (0.5)	0 (0.0)	0 (0.0)
missing	0 (0.0)	1 (0.2)	1 (0.2)	2 (0.5)
Total	402	402	402	402

Note. The number in the parenthesis represents the percentage associated with that frequency count.

Table 3.
Frequency Distribution Based On the Original Response Categories for the 3rd Data Set[a]

Response	Pre-test				Post-test			
Category	8	9	10	13	8	9	10	13
A	4 (2.0)	11 (5.4)	33 (16.1)	40 (19.5)	63 (32.0)	68 (34.5)	83 (42.1)	111 (56.3)
B	3 (1.5)	4 (2.0)	138 (67.3)	140 (68.3)	11 (5.6)	8 (4.1)	98 (49.7)	69 (35.0)
C	3 (1.5)	1 (0.5)	25 (12.2)	12 (5.9)	6 (3.0)	5 (2.5)	8 (4.1)	10 (5.1)
D	3 (1.5)	175 (85.4)	5 (2.4)	2 (1.0)	7 (3.6)	109 (55.3)	7 (3.6)	3 (1.5)
E	18 (8.8)	8 (3.9)	1 (0.5)	2 (1.0)	15 (7.6)	3 (1.5)	1 (0.5)	1 (0.5)
F	60 (29.3)	3 (1.5)	0 (0.0)	5 (2.4)	19 (9.6)	1 (0.5)	0 (0.0)	2 (1.0)
G	111 (54.1)	0 (0.0)	2 (1.0)	0 (0.0)	76 (38.6)	3 (1.5)	0 (0.0)	0 (0.0)
J	1 (0.5)	3 (1.5)	0 (0.0)	1 (0.5)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.5)
missing	2 (1.0)	0 (0.0)	1 (0.5)	3 (1.5)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Total	205	205	205	205	197	197	197	197

Table 3 (continued).
Frequency Distribution Based On the Original Response Categories for the 3rd Data Set[a]

Response Category	Combined			
	8	9	10	13
A	67 (16.7)	79 (19.7)	116 (28.9)	151 (37.6)
B	14 (3.5)	12 (3.0)	236 (58.7)	209 (52.0)
C	9 (2.2)	6 (1.5)	33 (8.2)	22 (5.5)
D	10 (2.5)	284 (70.6)	12 (3.0)	5 (1.2)
E	33 (8.2)	11 (2.7)	2 (0.5)	3 (0.7)
F	79 (19.7)	4 (1.0)	0 (0.0)	7 (1.7)
G	187 (46.5)	3 (0.7)	2 (0.5)	0 (0.0)
J	1 (0.5)	3 (0.7)	0 (0.0)	2 (0.5)
missing	2 (1.0)	0 (0.0)	0 (0.0)	3 (0.7)
Total	402	402	402	402

Note. The number in the parenthesis represents the percentage associated with that frequency count.
[a]Questions 8, 9, 10, and 13 only – the frequency distribution for the other four questions (2, 5, 11, and 12) can be found in Table 2.

Table 4.
Frequency Distribution Based On the Original Response Categories for the 4th Data Set

Response	Pre-test					Post-test				
Category	22	23	24	25	26	22	23	24	25	26
A	55 (27.2)	5 (2.5)	23 (11.3)	8 (3.9)	117 (57.6)	140 (71.4)	12 (6.1)	2 (1.0)	23 (11.7)	35 (17.9)
B	0 (0.0)	27 (13.3)	94 (46.3)	25 (12.3)	10 (4.9)	6 (3.1)	114 (58.2)	34 (17.3)	105 (53.6)	4 (2.0)
C	1 (0.5)	2 (1.0)	47 (23.2)	2 (1.0)	50 (24.6)	4 (2.0)	4 (2.0)	144 (73.5)	8 (4.1)	149 (76.0)
D	1 (0.5)	15 (7.4)	2 (1.0)	4 (2.0)	0 (0.0)	0 (0.0)	3 (1.5)	2 (1.0)	0 (0.0)	1 (0.5)
E	139 (68.5)	3 (1.5)	2 (1.0)	25 (12.3)	13 (6.4)	46 (23.5)	0 (0.0)	0 (0.0)	7 (3.6)	4 (2.0)
F	1 (0.5)	24 (11.8)	14 (6.9)	95 (46.8)	0 (0.0)	0 (0.0)	7 (3.6)	7 (3.6)	41 (20.9)	0 (0.0)
G	4 (2.0)	113 (55.7)	8 (3.9)	18 (8.9)	6 (3.0)	0 (0.0)	45 (23.0)	4 (2.0)	4 (2.0)	2 (1.0)
J	2 (1.0)	13 (6.4)	11 (5.4)	23 (11.3)	6 (3.0)	0 (0.0)	10 (5.1)	2 (1.0)	7 (3.6)	0 (0.0)
missing	0 (0.0)	1 (0.5)	2 (1.0)	3 (1.5)	1 (0.5)	0 (0.0)	1 (0.5)	1 (0.5)	1 (0.5)	1 (0.5)
Total	203	203	203	203	203	196	196	196	196	196

Table 4 (continued).
Frequency Distribution Based On the Original Response Categories for the 4th Data Set

Response Category	Combined				
	22	23	24	25	26
A	195 (48.9)	17 (4.3)	25 (6.3)	31 (7.8)	152 (38.1)
B	6 (1.5)	141 (35.3)	128 (32.1)	130 (32.6)	14 (3.5)
C	5 (1.3)	6 (1.5)	191 (47.9)	10 (2.5)	199 (49.9)
D	1 (0.3)	18 (4.5)	4 (1.0)	4 (1.0)	1 (0.3)
E	185 (46.4)	3 (0.8)	2 (0.5)	32 (8.0)	17 (4.3)
F	1 (0.3)	31 (7.8)	21 (5.3)	136 (34.1)	0 (0.0)
G	4 (1.0)	158 (39.6)	12 (3.0)	22 (5.5)	8 (2.0)
J	2 (0.5)	23 (5.8)	13 (3.3)	30 (7.5)	6 (1.5)
missing	0 (0.0)	2 (0.5)	3 (0.8)	4 (1.0)	2 (0.5)
Total	399	399	399	399	399

Note. The number in the parenthesis represents the percentage associated with that frequency count.

Table 5.
Frequency Distribution Based On the Three Response Categories for the 1st Data Set

Response	Pre-test					Post-test				
Category	5	9	18	22	28	5	9	18	22	28
1	19 (19.2)	12 (12.1)	14 (14.1)	56 (56.6)	6 (6.1)	28 (28.6)	21 (21.4)	55 (56.1)	43 (43.9)	7 (7.1)
2	51 (51.5)	48 (48.5)	47 (47.5)	39 (39.4)	85 (85.9)	63 (64.3)	41 (41.8)	15 (15.3)	54 (55.1)	87 (88.8)
3	29 (29.3)	39 (39.4)	37 (37.4)	N/A[a]	3 (3.0)	7 (7.1)	36 (36.7)	28 (28.6)	N/A[a]	3 (3.1)
missing	0 (0.0)	0 (0.0)	1 (1.0)	4 (4.0)	5 (5.1)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.0)	1 (1.0)
Total	99	99	99	99	99	98	98	98	98	98

Table 5 (continued).
Frequency Distribution Based On the Three Response Categories for the 1st Data Set

Response	Combined				
Category	5	9	18	22	28
1	47 (23.9)	33 (16.8)	69 (35.0)	99 (50.3)	13 (6.6)
2	114 (57.9)	89 (45.2)	62 (31.5)	93 (47.2)	172 (87.3)
3	36 (18.3)	75 (38.1)	65 (33.0)	N/A[a]	6 (3.0)
missing	0 (0.0)	0 (0.0)	1 (0.5)	5 (2.5)	6 (3.0)
Total	197	197	197	197	197

Note. The number in the parenthesis represents the percentage associated with that frequency count.
[a]There is no response category 3 for item 22 after recoding the data based on the three-response coding scheme.

Table 6.
Frequency Distribution Based On the Three Response Categories for the 2nd Data Set

Response	Pre-test				Post-test			
Category	2	5	11	12	2	5	11	12
1	13 (6.3)	71 (34.6)	13 (6.3)	10 (4.9)	79 (40.1)	126 (64.0)	96 (48.7)	94 (47.7)
2	179 (87.3)	94 (45.9)	125 (61.0)	176 (85.9)	102 (51.8)	51 (25.9)	55 (27.9)	84 (42.6)
3	13 (6.3)	39 (19.0)	66 (32.2)	17 (8.3)	16 (8.1)	20 (10.2)	46 (23.4)	19 (9.6)
missing	0 (0.0)	1 (0.5)	1 (0.5)	2 (1.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Total	205	205	205	205	197	197	197	197

Table 6 (continued).
Frequency Distribution Based On the Three Response Categories for the 2nd Data Set

Response	Combined			
Category	2	5	11	12
1	92 (22.9)	197 (49.0)	109 (27.1)	104 (25.9)
2	281 (69.9)	145 (36.1)	180 (44.8)	260 (64.7)
3	29 (7.2)	59 (14.7)	112 (27.9)	36 (9.0)
missing	0 (0.0)	1 (0.2)	1 (0.2)	2 (0.5)
Total	402	402	402	402

Note. The number in the parenthesis represents the percentage associated with that frequency count.

Table 7.
Frequency Distribution Based On the Three Response Categories for the 3rd Data Set[a]

Response	Pre-test				Post-test			
Category	8	9	10	13	8	9	10	13
1	4 (2.0)	11 (5.4)	33 (16.1)	40 (19.5)	63 (32.0)	68 (34.5)	83 (42.1)	111 (56.3)
2	111 (54.1)	175 (85.4)	138 (67.3)	140 (68.3)	76 (38.6)	109 (55.3)	98 (49.7)	69 (35.0)
3	88 (42.9)	19 (9.3)	33 (16.1)	22 (10.7)	58 (29.4)	20 (10.2)	16 (8.1)	17 (8.6)
missing	2 (1.0)	0 (0.0)	1 (0.5)	3 (1.5)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Total	205	205	205	205	197	197	197	197

Table 7 (continued).
Frequency Distribution Based On the Three Response Categories for the 3rd Data Set[a]

Response	Combined			
Category	8	9	10	13
1	67 (16.7)	79 (19.7)	116 (28.9)	151 (37.6)
2	187 (46.5)	284 (70.6)	236 (58.7)	209 (52.0)
3	146 (36.3)	39 (9.7)	49 (12.2)	39 (9.7)
missing	2 (1.0)	0 (0.0)	0 (0.0)	3 (0.7)
Total	402	402	402	402

Note. The number in the parenthesis represents the percentage associated with that frequency count.
[a]Questions 8, 9, 10, and 13 only – the frequency distribution for the other four questions (2, 5, 11, and 12) can be found in Table 6.

Table 8.
Frequency Distribution Based On the Three Response Categories for the 4th Data Set

Response	Pre-test					Post-test				
Category	22	23	24	25	26	22	23	24	25	26
1	55 (27.1)	32 (15.8)	47 (23.2)	33 (16.3)	50 (24.6)	146 (74.5)	126 (64.3)	144 (73.5)	128 (65.3)	149 (76.0)
2	140 (69.0)	137 (67.5)	117 (57.6)	120 (59.1)	127 (62.6)	46 (23.5)	52 (26.5)	36 (18.4)	48 (24.5)	39 (19.9)
3	8 (3.9)	33 (16.3)	37 (18.2)	47 (23.2)	25 (12.3)	4 (2.0)	17 (8.7)	15 (7.7)	19 (9.7)	7 (3.6)
missing	0 (0.0)	1 (0.5)	2 (1.0)	3 (1.5)	1 (0.5)	0 (0.0)	1 (0.5)	1 (0.5)	1 (0.5)	1 (0.5)
Total	203	203	203	203	203	196	196	196	196	196

Table 8 (continued).
Frequency Distribution Based On the Three Response Categories for the 4th Data Set

Response	Combined				
Category	22	23	24	25	26
1	201 (50.4)	158 (39.6)	191 (47.9)	161 (40.4)	199 (49.9)
2	186 (46.6)	189 (47.4)	153 (38.3)	168 (42.1)	166 (41.6)
3	12 (3.0)	50 (12.5)	52 (13.0)	66 (16.5)	32 (8.0)
missing	0 (0.0)	2 (0.5)	3 (0.8)	4 (1.0)	2 (0.5)
Total	399	399	399	399	399

Note. The number in the parenthesis represents the percentage associated with that frequency count.

Table 9.
The Pearson Correlations for the 1st Data Set

	5	9	18	22	28
			Pre-test		
5	1.000				
9	-.016	1.000			
18	-.051	-.030	1.000		
22	-.139	-.034	.012	1.000	
28	-.032	.043	.032	.156	1.000
test	.462	.494	.511	.323	.327
			Post-test		
5	1.000				
9	-.068	1.000			
18	-.246	-.013	1.000		
22	.174	.089	-.130	1.000	
28	.183	.029	-.039	.016	1.000
test	.355	.558	.491	.405	.309
			Combined		
5	1.000				
9	-.018	1.000			
18	-.059	.007	1.000		
22	-.034	.017	-.108	1.000	
28	.067	.036	-.004	.080	1.000
test	.454	.528	.539	.300	.308

Table 10.
The Pearson Correlations for the 2nd Data Set

	2	5	11	12
		Pre-test		
2	1.000			
5	.191	1.000		
11	.219	.133	1.000	
12	.227	.133	.387	1.000
test	.546	.703	.677	.581
		Post-test		
2	1.000			
5	.320	1.000		
11	.193	.169	1.000	
12	.252	.227	.536	1.000
test	.614	.614	.742	.743
		Combined		
2	1.000			
5	.317	1.000		
11	.284	.228	1.000	
12	.329	.262	.555	1.000
test	.640	.669	.759	.742

Table 11.
The Pearson Correlations for the 3rd Data Set

	2	5	11	12	8	9	10	13
				Pre-test				
2	1.000							
5	.191	1.000						
11	.219	.133	1.000					
12	.227	.133	.387	1.000				
8	.130	.148	.475	.207	1.000			
9	.216	.058	.225	.416	.259	1.000		
10	.217	.083	.158	.121	.130	.068	1.000	
13	.151	.116	.266	.270	.128	.111	.274	1.000
test	.488	.516	.659	.560	.577	.469	.491	.558
				Post-test				
2	1.000							
5	.320	1.000						
11	.193	.169	1.000					
12	.252	.227	.536	1.000				
8	.120	.080	.518	.417	1.000			
9	.128	.075	.300	.607	.581	1.000		
10	.180	.072	.463	.343	.639	.494	1.000	
13	.227	.249	.541	.456	.455	.315	.528	1.000
test	.447	.414	.740	.735	.749	.665	.713	.726
				Combined				
2	1.000							
5	.317	1.000						
11	.284	.228	1.000					
12	.329	.262	.555	1.000				
8	.206	.181	.557	.430	1.000			
9	.220	.131	.342	.597	.526	1.000		
10	.256	.144	.401	.332	.483	.380	1.000	
13	.274	.251	.496	.456	.397	.303	.466	1.000
test	.527	.502	.754	.738	.732	.644	.657	.699

Table 12.
The Pearson Correlations for the 4th Data Set

	22	23	24	25	26
			Pre-test		
22	1.000				
23	.545	1.000			
24	.348	.410	1.000		
25	.523	.465	.348	1.000	
26	.482	.354	.642	.430	1.000
test	.742	.735	.748	.742	.781
			Post-test		
22	1.000				
23	.608	1.000			
24	.301	.377	1.000		
25	.584	.571	.388	1.000	
26	.268	.372	.644	.366	1.000
test	.725	.797	.720	.789	.705
			Combined		
22	1.000				
23	.655	1.000			
24	.455	.506	1.000		
25	.641	.614	.489	1.000	
26	.513	.488	.716	.523	1.000
test	.794	.815	.795	.816	.809

Table 13.
The Polyserial Correlations between the Items and the Test for the 1st Data Set

5	9	18	22	28
		Pre-test		
.515	.559	.575	.409	.509
		Post-test		
.419	.624	.586	.510	.473
		Combined		
.510	.593	.607	.376	.476

Table 14.
The Polyserial Correlations between the Items and the Test for the 2nd Data Set

2	5	11	12
	Pres-test		
.777	.785	.791	.831
	Post-test		
.704	.748	.852	.855
	Combined		
.764	.767	.843	.864

Table 15.
The Polyserial Correlations between the Items and the Test for the 3rd Data Set

2	5	11	12	8	9	10	13
			Pre-test				
.693	.576	.769	.797	.692	.643	.572	.656
			Post-test				
.513	.505	.850	.846	.835	.758	.819	.855
			Combined				
.627	.574	.838	.859	.820	.762	.748	.797

Table 16.
The Polyserial Correlations between the Items and the Test for the 4th Data Set

22	23	24	25	26
		Pre-test		
.915	.857	.843	.839	.897
		Post-test		
.960	.977	.945	.972	.945
		Combined		
.948	.923	.910	.920	.939

BUGS Analyses

The model-based analysis procedure described in the previous chapter was followed to analyze each data set under three models, i.e., the homogeneous, partially homogeneous, and heterogeneous AR models. The *DIC* was used to choose a model that best fits the data, so our discussion below is based on the preferred model given that data set. Since we analyzed four different data sets, namely, the FCI data with 5 items, the Force-Motion data with 4 items, the Force-Motion data with 8 items (4 of them are the same as the second data set), and the Acceleration data with 5 items, we present the results in following manner:

First, since these data sets are repeatedly mentioned here, it is convenient to name each data set FCI5, FM4, FM8, and Acc5, respectively. Second, each of them is discussed separately except for FM4 and FM8. Because these two analyses concern an overlapping set of items, it is of interest to compare the item parameter estimates between these two for common items. Next, we interpret both item and person parameters in detail using FCI5 only since the interpretation would be analogous for the remaining data sets. Finally, for each data set, we focus on item level analysis (e.g., patterns of changes that occurred from pre to post and what kinds of learning might have taken place), so each item is discussed thoroughly. In terms of persons, however, the tests are so short it is unproductive to discuss each examinee's performance or change over time. Instead, we examine the overall students' propensity level distributions on both pre- and post-test, as it relates to the tendencies of change for performance on items.

FCI5. The *DIC* values for the heterogeneous, partially homogeneous, and homogeneous models for the FCI5 data set are 1730.170, 1785.390, and 1780.340 respectively. Thus the heterogeneous model is preferred, indicating that both the conditional probabilities for items and the population distributions change over time. To further examine the nature of the performance patterns, we will discuss the item parameter estimates that are summarized in Table 17.

We first demonstrate how to interpret the item parameter under the AR model using the items on the pre-test; that is, b[1,j,k] (from now on, we refer as items 1 through 5 the questions numbered 5, 9, 18, 22, and 28, respectively, in the full FCI assessment). From Table 17, we can see that items 1, 2, and 3 on the pre-test have a greater tendency for eliciting responses using model 3, the null model, since b[1,j,3] is the greatest among three vector elements (0.2798, 0.7817, and 0.6731 for b[1,1,3], b[1,2,3], and b[1,3,3], respectively), indicating that Aristotelian responses, and presumably a nonscientific way of thinking, are more common on these items, given all other things being equal. Item 4 tends to evoke responses based on model 1, the expert model, using the Newtonian approach, as indicated by the fact that b[1,4,1] is greater than other two vector elements (2.1230 vs. 1.0150 and -3.1370). Recall that having modeled the data as if response category 3 exists for item 4 but was not chosen, we can see that b[1,4,3] is extremely low (-3.1370) as would be expected. Item 5, on the other hand, has a greater tendency to evoke responses using model 2, an incorrect student model (1.9310 vs. -0.4903 and -1.4410), indicating that "impetus theory" belief is more common on this item, again given all other things being equal.

Table 17.
The Item Parameter Estimates for the FCI Data under the Heterogeneous AR Model

Item Parameter	Mean	SD	MC error
(Pre-test)			
b[1,1,1][a]	-0.4088	0.7457	0.01978
b[1,1,2]	0.1290	0.7439	0.02174
b[1,1,3]	0.2798	N/A[b]	N/A
b[1,2,1]	-0.8745	0.7524	0.01977
b[1,2,2]	0.0928	0.7445	0.02173
b[1,2,3]	0.7817	N/A	N/A
b[1,3,1]	-0.7284	0.7496	0.01976
b[1,3,2]	0.0553	0.7448	0.02175
b[1,3,3]	0.6731	N/A	N/A
b[1,4,1]	2.1230	0.8096	0.01966
b[1,4,2]	1.0150	0.8144	0.02152
b[1,4,3]	-3.1370	N/A	N/A
b[1,5,1]	-0.4903	0.7863	0.01950
b[1,5,2]	1.9310	0.7670	0.02155
b[1,5,3]	-1.4410	N/A	N/A
(Post-test)			
b[2,1,1]	0.1814	0.7706	0.02142
b[2,1,2]	0.8759	0.7591	0.02093
b[2,1,3]	-1.0570	N/A	N/A
b[2,2,1]	-0.6220	0.7672	0.02148
b[2,2,2]	-0.1211	0.7540	0.02099
b[2,2,3]	0.7431	N/A	N/A
b[2,3,1]	0.6328	0.7625	0.02150
b[2,3,2]	-1.0900	0.7629	0.02093
b[2,3,3]	0.4573	N/A	N/A
b[2,4,1]	1.5500	0.8373	0.02132
b[2,4,2]	1.5450	0.8277	0.02077
b[2,4,3]	-3.0950	N/A	N/A
b[2,5,1]	-0.5777	0.8013	0.02114
b[2,5,2]	1.9620	0.7749	0.02079
b[2,5,3]	-1.3840	N/A	N/A

Note. The number of MC draws (i.e., the sample size) that were used to compute these statistics is 210,000.
[a]b[t,j,k] represents the parameter estimate at the time point *t* (t = 1 for the pre-test while t =2 for the post-test) for the item *j* with the response category *k*.
[b]Since the parameters with response category 3 are not estimated here – they were obtained by computing the sum of parameter estimates with the first two response categories and then reversing the sign of the sum, the associated SDs and MC errors are not available.

Comparisons within and between items are possible based on those parameter estimates (although we do not intend to do it here in detail). In terms of the significance of differences between parameter estimates within items, one can compare the posterior means using an independent or dependent t-test (which test is appropriate depends on the correlation of the estimates of selected parameters, and the correlation can be obtained through *WinBUGS* under **inference** menu). The (posterior) standard deviations of the parameter estimates (labeled SD as listed in the table) can be used much as standard errors of estimates that are (often improperly) used in frequentist statistical inference. However, given the fact that FCI5 only contains 197 students' responses on 5 items, the posterior standard deviations for the individual parameters in these analyses are relatively large and comparisons may not be reliable (this is also true for the remaining data sets). Therefore, the word "tendency" is used here (also for other data sets as well) mainly to describe the patterns shown from the parameter estimates; statistically significant differences among population values are not claimed. If one desired more accurate comparisons, larger samples of students (along with preferably longer tests) are required to make strong inference about item parameters. On the other hand, these parameter estimates are indeed descriptive of patterns in the sample, which is in fact nearly equivalent to the population of interest in the classes the FCI5 is from. On the other hand, these data were sufficient to test for and find differences among student models as whole, as we compared three nested models under the AR model.

Since the AR model belongs to the Rasch family, one would be able to compare two items in such a way that only the parameters associated with those two items are involved in the comparison (this is called "specific objectivity", a unique feature of Rasch

models). In the context of the AR model, this means that we can compare two items for a given dimension (Newtonian strategy, "impetus theory" belief, or Aristotelian thinking). One way of doing it is to use the log of the odds ratio (or logit). For example, we can compare items 1 and 2 (use Table 17) in terms of the Newtonian dimension. The difference between b[1,1,1] and b[1,2,1] is 0.4657 (i.e., b[1,1,1] minus b[1,2,1]). This, in fact, indicates the difference is 0.4657 in a logit metric (i.e., 0.4657 logits). We then can transform this logit to odds, i.e., log odds = 0.4657, so odds = exp(0.4657) = 1.6. This means that item 1 is more likely than item 2 to provoke a Newtonian response with the odds of 1.6:1. Similar comparisons can be made across persons, to compare their relative propensities to give responses from the various classes.

By looking at the item text, we may gain further insight into why each item has a different tendency to provoke responses based on different physics models. This is discussed below.

Item 1 asks what kind of force(s) is (are) acting on the ball when it leaves a boy's hand (a boy throws it straight up) and later returns to the ground. One of choices states that "the ball falls back down to the earth simply because that is its natural action." This is based on the Aristotelian way of thinking, and would be a choice for naïve students without knowledge of Newton's laws since it is close to everyday thinking. Similarly, item 2 asks the same kind of question as item 1 and provides a choice also based on the naïve thinking: "gravity does not exert a force on the puck; it falls because of intrinsic tendency of the object to fall to its natural place."

The scenario for item 3 is a little more complicated than the first two items since it involves an elevator being lifted up an elevator shaft by a steel cable and asks how the upward force by the cable and the downward force due to gravity act on the elevator when it is moving up the shaft at a constant velocity. The keyword here is "constant" that is being highlighted in the question, implying that those two forces are equal. If students use a Newtonian approach, which is unlikely during the pre-test, they would choose the correct answer. Even for those students who possess some knowledge about Newtonian mechanics, they would select choices (if not the correct one) stating that the upward force is greater than the downward force for the elevator is moving up but without considering whether its velocity is constant. However, for students who answer without knowing anything about Newton's laws for force-motion relationships or simply guess the answer, the choices corresponding to model 3 would be selected – "it goes up because the cable is being shortened, not because of the force being exerted on the elevator by the cable" (naïve thinking) or "the upward force on the elevator by the cable is less than the downward force of gravity" (the unsystematic approach), especially on the pre-test. Therefore, on the pre-test item 3 has a greater tendency to provoke thinking that leads to a response classified as model 3.

Item 4 seems to be an easy one since the question simply asks what kinds of force(s) – the force of gravity, the force of the "hit", or the force of air resistance – is (are) acting on the golf ball during its entire flight after it was hit. Unlike the first two items, it does not ask the sign or magnitude of the force(s) acting on the ball; that is, students do not need to justify how the forces affect the ball's movement. In this sense, item 4 is much simpler than the first two questions. Without knowing much about force-motion relation

or simply using common sense, students would be more likely to select a correct choice – the one stating the force of gravity and the one consisting of both the force of gravity and force of air resistance. This may be why item 4 tends to provoke a response consistent with the correct Newtonian model on the pre-test data. Since other choices also include the force of "hit", students who select those choices would be thought of as using a model 2 approach, an incorrect model, for the force of "hit" no longer acts on the ball after it was hit. Note that there are no choices provided based on naïve thinking as for the first two items – i.e., no responses that would be coded as using model 3.

For item 5, the situation is similar to item 3, but it involves friction forces. A large box is being pushed across the floor at a constant speed of 4.0 m/s, and the question asks how the forces are acting on the box. Again, the keyword here is "constant" as for item 3. The item would be answered correctly if a student were using a Newtonian way of thinking. However, some of the choices accompanying this item make it harder than it would be to respond in this way. For example, the first choice states that "If the force applied to the box is doubled, the constant speed of the box will increase to 8.0 m/s." This statement is not true even without considering the frictional forces since velocity is the function of time (i.e., the acceleration is constant but the speed of the box is not constant). One other hand, the frictional forces do exist in most real-world situations, and the net force should be determined before we can figure out the speed of the box. Students could pick this choice mainly because they thought the double force will result in a constant acceleration (without considering the frictional forces), which in turns yields a constant double speed (without fully understanding the relationship between velocity and acceleration). Another plausible choice is the last one, stating that "There is a force being

applied to the box to make it move but the external forces such as friction are not real forces they just resist motion." This statement looks correct, but friction is real even when the box is at motion. The reason the box being pushed is moving at a constant speed (implying the acceleration is zero) is that the force applied to the box equals the frictional force (so the net force is zero). Students who select this choice would be considered as using model 2 approach, an incorrect understanding about forces or force-motion relationships. Thus, these two choices would make item 5 have a greater tendency to evoke a response using an incorrect student model.

In terms of person parameter estimates, we use the first five examinees' parameter estimates from the pre-test, labeled theta[1,i,k] (see Table 18), to demonstrate how to interpret the parameter estimates under the AR model. Recall that each student's propensity parameters are constrained to sum to zero, so the parameter estimates have an ipsative quality: A student has propensities for each class of response, but since some response must be made, the comparison is among how likely they are within the examinee (again other things being equal). As seen in Table 18, student 1 has a greater propensity to use the model 3, given all other things being equal, for problem-solving since the theta[1,1,3] (0.2869) is greater than theta[1,1,1] (-0.3883) and theta[1,1,2] (0.1014). Student 2 is inclined to use both models 1 and 3 since theta[1,2,1] (0.1507) and theta[1,2,3] (0.2023) are greater than theta[1,2,2] (-0.3529) and they are slightly different, indicating that he/she inconsistently uses different models for problem-solving. Depending on features of items, such a student provides some Newtonian responses, some "impetus-theory" responses, and some nonscientific responses, with propensities suggested by the θ estimates. Similarly, students 3 and 5 are also in a mixed model state

for they tend to use more than one model to respond to physics questions. The first of these two students has a greater propensity to use either model 2 or 3, and the latter has a greater propensity to use either model 1 or 2. Student 4, on the other hand, is closer to being in a "pure" model state, as was the case for student 1; student 4 has a much greater propensity to use model 1 (theta[1,4,1] is much greater than both theta[1,4,2] and theta[1,4,3]), indicating that he/she is inclined to use Newtonian approach for solving physics tasks like those in this set of FCI items.

After instruction, students' propensity to use the various models may change. For example, on the post-test student 1 (as labeled theta[2,100,k] in the table) now has a greater propensity to use model 2 (on the pre-test he/she tends to use model 3), indicating that he/she may have gained some knowledge about force-motion relation (but not fully reached the point of using it consistently). Such individual changes in turn reflect on the item's tendency to evoke a certain model on the post-test. This is discussed below.

Table 18.
The First 5 Examinees' Parameter Estimates for the FCI Data under the Heterogeneous AR Model

Person Parameter	Mean	SD	MC error
(Pre-test)			
theta[1,1,1][a]	-0.3883	1.0160	0.02167
theta[1,1,2]	0.1014	0.9380	0.02269
theta[1,1,3]	0.2869	N/A[b]	N/A
theta[1,2,1]	0.1507	0.9821	0.02166
theta[1,2,2]	-0.3529	0.9541	0.02262
theta[1,2,3]	0.2023	N/A	N/A
theta[1,3,1]	-1.0100	1.0540	0.02166
theta[1,3,2]	0.5528	0.9348	0.02266
theta[1,3,3]	0.4572	N/A	N/A
theta[1,4,1]	1.1480	1.0070	0.02160
theta[1,4,2]	0.5642	0.9874	0.02263
theta[1,4,3]	-1.7120	N/A	N/A
theta[1,5,1]	0.2422	1.0330	0.02165
theta[1,5,2]	0.2905	0.9771	0.02265
theta[1,5,3]	-0.5327	N/A	N/A

Table 18 (continued).
The First 5 Examinee's Parameter Estimates for the FCI Data under the Heterogeneous AR Model

Person Parameter	Mean	SD	MC error
(Post-test)			
theta[2,1,1]	0.7119	1.0120	0.02154
theta[2,1,2]	0.9340	1.0080	0.02077
theta[2,1,3]	-1.6460	N/A	N/A
theta[2,2,1]	0.7068	1.0170	0.02165
theta[2,2,2]	0.9316	1.0090	0.02089
theta[2,2,3]	-1.6380	N/A	N/A
theta[2,3,1]	0.2805	0.9605	0.02170
theta[2,3,2]	0.0821	0.9611	0.02068
theta[2,3,3]	-0.3626	N/A	N/A
theta[2,4,1]	1.0900	1.0060	0.02171
theta[2,4,2]	0.5356	1.0110	0.02076
theta[2,4,3]	-1.6250	N/A	N/A
theta[2,5,1]	0.7092	1.0130	0.02162
theta[2,5,2]	0.9267	1.0070	0.02082
theta[2,5,3]	-1.6360	N/A	N/A

Note. The number of MC draws used to compute these statistics is 210,000.
[a]theta[t,i,k] represents the parameter estimate at the time point *t* (t = 1 for the pre-test while t =2 for the post-test) for the person *i* with the response category *k*.
[b]Since the parameters with response category 3 are not estimated here – they were obtained by computing the sum of parameter estimates with the first two response categories and then reversing the sign of the sum, the associated SDs and MC errors are not available.

Table 19 summarizes the average parameter estimates over persons and items for the pre- and post-test. We can see that on average examinees tend to use model 2 on the pre-test (labeled as mu12). After instruction, they still have a greater propensity to use model 2 (labels as mu22). However, their tendency to use model 1 has been increased (-0.0947 on the pre vs. 0.2144 on the post), while the tendency to use model 3 has been decreased (-0.4938 on the pre vs. -0.7124 on the post). The average tendency over items to elicit certain models (labeled as mub*tk*) also shows the similar result. This indicates that even though on average students tend not to use Newtonian approach for problem-solving after instruction, some improvement has occurred – students are in a transition toward understanding Newtonian mechanics. On the other hand, this also implies that students

have difficulties in understanding some concepts. Just which ones may be examined by looking at the change from individual items.

To conveniently study the change for the item's tendency to provoke specific responses based on different student models, we represent Table 17 in terms of what model is likely to be elicited given the features of the item, as seen in Table 20.

Table 19.
The Average Parameter Estimates over Persons and Items before and after Instruction for the FCI Data

Parameters	Mean	SD	MC error
(Person)			
mu11[a]	-0.0947	0.7611	0.02145
mu12	0.5885	0.7467	0.02243
mu13	-0.4938	N/A[b]	N/A
mu21	0.2144	0.7483	0.02145
mu22	0.4980	0.7371	0.02055
mu23	-0.7124	N/A	N/A
(Item)			
mub11[a]	-0.0609	0.7549	0.01795
mub12	0.5618	0.7452	0.01878
mub13	-0.5009	N/A	N/A
mub21	0.2023	0.7465	0.01797
mub22	0.5563	0.7395	0.01727
mub23	-0.7586	N/A	N/A

Note. The number of MC draws used to compute these statistics is 210,000.
[a]mu*tk* represents the parameter estimate at the time point *t* (t = 1 for the pre-test while t =2 for the post-test) over persons with response category *k*. Similarly, mub*tk* represents the parameter estimate at the time point *t* (t = 1 for the pre-test while t =2 for the post-test) over items with response category *k*.
[b]Since the parameters with response category 3 are not estimated here – they were obtained by computing the sum of parameter estimates with the first two response categories and then reversing the sign of the sum, the associated SDs and MC errors are not available.

Table 20.
The Model Elicited Given the Features of the Item before and after Instruction for the FCI Data

	1 (5)[a]	2 (9)	3(18)	4(22)	5(28)
Before	model 3	model 3	Model 3	model 1	model 2
After	model 2	model 3	Model 1	model 1 or 2	model 2

[a]The number in the parenthesis refers to the original item number.

From Table 20, we can see that after instruction the modal tendency to evoke the response based on a certain model has changed for 3 items (out of 5 items), namely items 1, 3, and 4. Item 1 has a greater tendency to evoke a response based on model 2 on the post-test (vs. model 3 on the pre-test). Item 3 tends to provoke a response based on model 1 on the post-test (vs. model 3 on the pre-test). Item 4, interestingly, elicits a response based on models 1 or 2 equally well on the post-test (vs. model 1 on the pre-test). There is no appreciable change across occasions for items 2 and 5.

First of all, this inconsistent change from item to item implies that although students' understanding about force (or force-motion relation) has been improved as discussed earlier, it needs not occur for everyone and needs not be constant across items. Even for items measuring the same concept from an expert's point of view, the students do not perform consistently. For example, regarding Newton's first law – an object can move with or without force – as represented by items 1, 2, and 4, students' responses on the post-test are not consistent, as indicated by different modal models elicited by these items.

Second, the results highlight how some concepts measured by these items are still difficult for students to comprehend after instruction. For example, Newton's first law as indicated above is one of them. Students still tend to believe that there is a force acting on the object to keep it moving as shown by item 1 (it tends to evoke a response based on

model 2) and even item 4 (it has a greater tendency to provoke a response based on either model 1 or 2). Notice that before instruction, item 4 tends to elicit the response based on model 1. It is surprising to see that kind of change for item 4 – perhaps implying that students become confused even after instruction. The concept about friction is another difficult one for students. Although students can use the correct model (i.e., constant speed implies equal force) to answer item 3, they do not apply the same approach to item 5. As discussed earlier, these two items ask the same kind of question, but item 5 additionally involves frictional forces. Since model 2 is the modal model for item 5 after instruction (the same as before instruction), this suggests that students still do not fully understand the frictional forces.

Comparison with Bao and Redish's analyses using FCI5. The five FCI questions used in the current study have been used by Bao and Redish, as mentioned before. Although the student population in their studies (mostly engineering majors) is quite different from the one in the current study (mostly biology majors), comparing their findings based on the methods they developed with the results derived from the AR analyses is of interest. Some discrepancies are expected since the data were obtained from two different populations and the statistics were derived from different methods of analyses.

In their study using *Concentration Analysis* (see Chapter II for more details), Bao and Redish (2001) pointed out that before instruction students tended to use common incorrect models (referred to as model 2 or 3 in this study since at that time the 3-model coding schema has not yet been developed) on those 5 FCI questions. This finding is similar to what we found here as seen in Table 20: Items have a greater tendency to elicit a

response based on model 2 or 3, except for item 4 which has a greater tendency to provoke the model 1 response. Note that the above finding from Bao and Redish's study is based on students in traditional classes (traditional lecturing plus sessions led by teaching assistants for recitations) as for the data used in the current study.

Bao and Redish's *Model Analysis* (2004) further indicated that students in the traditional classes tend to use model 2 (the incorrect model) before instruction, whereas they inconsistently use either model 1 or 2 (equally likely) after instruction. This is slightly different from what we found here as shown by Table 19 and discussed earlier. This may be due to different student data used and analyses employed in both studies. However, to some extents, the results from both analyses indicate that students are still having difficulties in understanding some force-motion concepts measured by FCI after instruction.

FMCE4 and FMCE8. As for FCI5, the heterogeneous model is preferred for both data sets. For the former, the *DIC* values for the homogeneous, partially homogeneous and heterogeneous models are 1785.390, 1780.340, and 1730.170, respectively; while for the latter, the corresponding *DIC* values for the models are 4595.720, 4494.520, and 4454.290, respectively. This indicates that both the item parameter estimates and the population distribution change after instruction for both data sets. To further understand this change and to compare the item parameter estimates between FMCE4 and FMCE8 for common items, we first discuss each data set separately, followed by the comparisons between these two data.

Table 21 summarizes the mean parameter estimates over persons and items before and after instruction for FMCE4. We can see that on average students tend to use model 2 before instruction but they have a greater propensity to use model 1 after instruction (although the difference between models 1 and 2 is not that substantial), indicating that most students have improved their understanding about force-motion concept. This shift is also seen by the change at the item level. On average, items tend to elicit a response based on model 2 before instruction, while they tend to evoke a response based on model 1 after instruction. We then summarize the parameter estimates for each item before and after instruction (Table 22), and represent the results in terms of the model that tends to be evoked before and after instruction (Table 23) to further study this change and its implications.

From Table 22 (or 23), we can see that before instruction items 1, 3, and 4 (as for FCI5, we refer items 1 through 4 to questions numbered 2, 5, 11, and 12, respectively) tend to provoke a response based on model 2, whereas item 2 has a greater tendency to evoke a response based on model 1. By looking at the item text, we may find clues to explain how this occurs.

Table 21.
The Average Parameter Estimates over Persons and Items before and after Instruction for the FMCE data with 4 Items

Parameters	Mean	SD	MC error
(Person)			
mu11[a]	-0.3941	0.7955	0.03276
mu12	0.6313	0.7003	0.02883
mu13	-0.2372	N/A[b]	N/A
mu21	0.4498	0.6971	0.02801
mu22	0.2872	0.7303	0.02968
mu23	-0.7370	N/A	N/A
(Item)			
mub11[a]	-0.3726	0.7786	0.02625
mub12	0.8668	0.7184	0.02328
mub13	-0.4942	N/A	N/A
mub21	0.3696	0.7135	0.02249
mub22	0.1622	0.7358	0.02392
mub23	-0.5318	N/A	N/A

Note. The number of MC draws used to compute these statistics is 100,000.
[a]mu*tk* represents the parameter estimate at the time point *t* (t = 1 for the pre-test while t =2 for the post-test) over persons with response category *k*. Similarly, mub*tk* represents the parameter estimate at the time point *t* (t = 1 for the pre-test while t =2 for the post-test) over items with response category *k*.
[b]Since the parameters with response category 3 are not estimated here – they were obtained by computing the sum of parameter estimates with the first two response categories and then reversing the sign of the sum, the associated SDs and MC errors are not available.

Table 22.
The Item Parameter Estimates for the FMCE Data with 4 Items under the Heterogeneous AR Model

Item Parameter	Mean	SD	MC error
(Pre-test)			
b[1,1,1][a]	-0.5376	0.8197	0.03261
b[1,1,2]	1.8090	0.7134	0.02880
b[1,1,3]	-1.2720	N/A[b]	N/A
b[1,2,1]	0.6444	0.8038	0.03287
b[1,2,2]	0.0341	0.7072	0.02891
b[1,2,3]	-0.6785	N/A	N/A
b[1,3,1]	-1.1470	0.8181	0.03280
b[1,3,2]	0.7052	0.7092	0.02890
b[1,3,3]	0.4416	N/A	N/A
b[1,4,1]	-0.8291	0.8237	0.03257
b[1,4,2]	1.7770	0.7137	0.02883
b[1,4,3]	-0.9482	N/A	N/A
(Post-test)			
b[2,1,1]	0.2585	0.7071	0.02803
b[2,1,2]	0.7819	0.7397	0.02977
b[2,1,3]	-1.0400	N/A	N/A
b[2,2,1]	0.9233	0.7065	0.02809
b[2,2,2]	-0.1167	0.7415	0.02976
b[2,2,3]	-0.8066	N/A	N/A
b[2,3,1]	0.2306	0.7042	0.02808
b[2,3,2]	-0.3001	0.7394	0.02977
b[2,3,3]	0.0695	N/A	N/A
b[2,4,1]	0.4393	0.7066	0.02808
b[2,4,2]	0.4625	0.7400	0.02979
b[2,4,3]	-0.9018	N/A	N/A

Note. The number of MC draws used to compute these statistics is 100,000.
[a]b[t,j,k] represents the parameter estimate at the time point *t* (t = 1 for the pre-test while t =2 for the post-test) for the item *j* with the response category *k*.
[b]Since the parameters with response category 3 are not estimated here – they were obtained by computing the sum of parameter estimates with the first two response categories and then reversing the sign of the sum, the associated SDs and MC errors are not available.

Table 23.
The Model Elicited Given the Features of the Item before and after Instruction for the FMCE Data with 4 items

	1 (2)[a]	2 (5)	3(11)	4(12)
Before	model 2	model 1	model 2	model 2
After	model 2	model 1	model 1	model 2 or 1

[a]The number in the parenthesis refers to the original item number.

The first two items are from the same set of questions involving a sled on ice. They ask which force (7 alternative choices as listed) would keep the sled moving in a certain velocity under the condition given by each question. Item 1 asks which force would keep the sled moving toward the right at a constant velocity. Students who answer without knowing and applying Newton's law might easily choose the force that is toward the right and is of constant strength (the choice B). However, this is based on an incorrect model (i.e., model 2) since, as mentioned earlier, an object can move with or without the force – the correct understanding based on Newton's first law. In other words, since the sled is moving there is no need to apply any force to keep it moving. Because of students' propensity to use "impetus theory" belief on this question before instruction, item 1 tends to elicit a response based on model 2.

Item 2, in fact, asks the same kind of question as item 1. However, it states the question in a different way. Instead of referring to a moving sled, it describes that "The sled was started from rest and pushed until it reached a steady (constant) velocity toward the right" and asks which force would keep the sled moving at this velocity. Without considering friction ("*Friction is so small that it can be ignored*" as mentioned in the beginning of this set of questions), it would be likely for students to select choice D: No applied force is needed, a response based on the correct model. This could be the reason why item 2 has a greater tendency to evoke a Newtonian or model 1 response even before instruction. It should be noted that, however, by selecting this correct choice does not indicate that students indeed know and apply Newtonian approach on this question. It is possible that students simply respond to this question based on their common sense, and it happens to be consistent with Newtonian way of thinking in this case. (This could be

further examined by interviewing students; i.e., by asking them to justify their reasons for choosing a specific choice) In addition, the tendency to elicit responses based on different models for the first 2 items (which are intrinsically equivalent) provides a piece of evidence about the context dependence discussed in Chapter II: students respond differently on two expert-equivalent situations due to the different features built into each scenario.

The next two items are also in the same set of questions. They involve a coin that is tossed straight up into the air, and they ask the direction and magnitude of the force acting on the coin in various cases. For item 3, the coin is moving upward after it is released; for item 4, the coin is at its highest point. Since students do not need to know exactly what force is acting on the coin (the accompanying choices imply there is only one type of force), they may respond to the question simply based on the position of the coin. Therefore, for item 3, students would tend to think the force that is up and decreasing because the coin is moving up (so the force is up), but it will stop moving up at some point (implying the force is decreasing). For item 4, it would be likely for them to believe the force is zero, since the coin being at its highest point implies no movement (no motion, so the magnitude of the force is zero). However, these rationales are based on the incorrect model (model 2): believing that there is a force acting on the coin, and its direction and magnitude would be changed depending on the position of the coin. Indeed, there is a force (and only one) acting on the coin after it is released: gravity, but it is always down and constant regardless the position of the coin. This is the correct model. Given the features of these items, the way of thinking characterized as model 2 appears to have been used by most students before instruction since it probably is consistent with their common sense.

Notice that the above scenario is similar to some of the questions on FCI5 – a boy throws a steel ball (question 5) or a golf ball is traveling through the air after hit (question 22). However, they are different in two ways. First, the questions on FMCE4 specifically describe the moving object (a coin in this case) at a certain position (moving upward or at the highest point), while on FCI5 the moving object (a steel ball or a golf ball) is described only in a very general way – it follows a parabolic-like path. Second, the accompanying choices for FMCE4 imply only one type of force acting on the coin, as mentioned above. Students do not need to know exactly what force(s) is (are) acting on the coin for FCME4, whereas the choices provided on FCI5 combine more features about the force(s) acting on the object – the type of force(s) and its (their) direction and magnitude. Because of these discrepancies, students (recall that FCI5 and FMCE4 data were obtained from similar populations) may respond differently – e.g., question 5 on FCI5 has a greater tendency to elicit a response based on model 3 (not model 2 as for FMCE4 questions) before instruction.

After instruction, the first two items retain their tendency to provoke certain models, while the last two items shift their tendencies to evoke different models. This suggests that students still have difficulties in understanding some concepts. First of all, students tend to believe that there is a force acting on an object to keep it moving – as shown by item 1 where both before and after instruction there is a greater tendency to observe a response based on model 2. This seems to be a common misconception, as suggested both here in FCI5 and as indicated in physics education literature. Second, students' understanding about gravity seems to be problematic (it has been improved but not fully) since students have a greater propensity to use either model 1 or 2 to respond to

item 4. They tend to think that the magnitude of the force (probably regardless of gravity or other types of force) is zero when the coin is at its highest point as discussed above.

Comparison of FMCE4 with Bao's analysis. These four questions on FMCE were also evaluated using *Model Analysis* in Bao's study (Bao, 1999). Bao indicated that most students in the traditional classes tend to use model 2 (the incorrect model) before instruction, while they tend to use either model 1 or 2 (more in model 2 but not by much) after instruction. This result is different from our analysis. As seen on Table 21 and discussed earlier, we found that students have a greater propensity to use model 2 on the pre-test, but they tend to use model 1 (rather than model 2) on the post-test. Again, this discrepancy could be due to the different student data used or to the different approaches of analyzing data.

Now we turn our discussion to FMCE8. Tables 24-26 list the average parameter estimates over persons and items before and after instruction, the item parameter estimates under the heterogeneous model, and the model elicited given the features of the item before and after Instruction for the FMCE8, respectively.

From Table 24, we can see that students have a greater propensity to use model 2 for problem-solving before instruction, whereas they tend to use either model 1 or 2 (equally likely) after instruction. The average tendency to elicit certain models over items also confirms this, although on average they tend to evoke the response based more on model 1 than model 2 after instruction. More detailed discussion about this change appears below. Regardless of this slight difference, the above result indicates that most students are in a mixed model state – in a transition toward understanding Newtonian mechanics.

Table 24.
The Average Parameter Estimates over Persons and Items before and after Instruction for the FMCE Data with 8 Items

Parameters	Mean	SD	MC error
(Person)			
mu11[a]	-0.5337	0.7199	0.03030
mu12	0.7374	0.6915	0.03004
mu13	-0.2073	N/A[b]	N/A
mu21	0.2609	0.7252	0.03100
mu22	0.2858	0.7470	0.03252
mu23	-0.5467	N/A	N/A
(Item)			
mub11[a]	-0.4007	0.7217	0.02709
mub12	0.8423	0.7002	0.02689
mub13	-0.4416	N/A	N/A
mub21	0.4205	0.7255	0.02770
mub22	0.2744	0.7438	0.02908
mub23	-0.6949	N/A	N/A

Note. The number of MC draws used to compute these statistics is 100,000.
[a]mu*tk* represents the parameter estimate at the time point *t* (t = 1 for the pre-test while t =2 for the post-test) over persons with response category *k*. Similarly, mub*tk* represents the parameter estimate at the time point *t* (t = 1 for the pre-test while t =2 for the post-test) over items with response category *k*.
[b]Since the parameters with response category 3 are not estimated here – they were obtained by computing the sum of parameter estimates with the first two response categories and then reversing the sign of the sum, the associated SDs and MC errors are not available.

Table 25.
The Item Parameter Estimates for the FMCE Data with 8 Items under the Heterogeneous AR Model

Item Parameter	Mean	SD	MC error
(Pre-test)			
b[1,1,1][a]	-0.4254	0.7515	0.03033
b[1,1,2]	1.6820	0.7071	0.03012
b[1,1,3]	-1.2570	N/A[b]	N/A
b[1,2,1]	0.7814	0.7286	0.03044
b[1,2,2]	-0.0940	0.6998	0.03017
b[1,2,3]	-0.6874	N/A	N/A
b[1,3,1]	-1.0110	0.7463	0.03038
b[1,3,2]	0.5946	0.7030	0.03020
b[1,3,3]	0.4164	N/A	N/A
b[1,4,1]	-0.7257	0.7566	0.03037
b[1,4,2]	1.6590	0.7076	0.03014
b[1,4,3]	-0.9331	N/A	N/A
b[1,5,1]	-1.9550	0.7804	0.03038
b[1,5,2]	0.7432	0.7118	0.03017
b[1,5,3]	1.2120	N/A	N/A
b[1,6,1]	-0.6990	0.7530	0.03032
b[1,6,2]	1.5670	0.7071	0.03018
b[1,6,3]	-0.8678	N/A	N/A
b[1,7,1]	0.0464	0.7339	0.03041
b[1,7,2]	0.6592	0.6998	0.03014
b[1,7,3]	-0.7055	N/A	N/A
b[1,8,1]	0.3744	0.7344	0.03046
b[1,8,2]	0.7675	0.7012	0.03015
b[1,8,3]	-1.1420	N/A	N/A
(Post-test)			
b[2,1,1]	0.4024	0.7365	0.03108
b[2,1,2]	0.7592	0.7584	0.03270
b[2,1,3]	-1.1620	N/A	N/A
b[2,2,1]	1.1260	0.7350	0.03112
b[2,2,2]	-0.1650	0.7589	0.03263
b[2,2,3]	-0.9615	N/A	N/A
b[2,3,1]	0.4115	0.7328	0.03110
b[2,3,2]	-0.3201	0.7566	0.03266
b[2,3,3]	-0.0914	N/A	N/A
b[2,4,1]	0.6058	0.7351	0.03109
b[2,4,2]	0.4310	0.7575	0.03267
b[2,4,3]	-1.0370	N/A	N/A

Table 25 (continued).
The Item Parameter Estimates for the FMCE Data with 8 Items under the Heterogeneous AR Model

Item Parameter	Mean	SD	MC error
b[2,5,1]	-0.2821	0.7351	0.03114
b[2,5,2]	0.0537	0.7552	0.03265
b[2,5,3]	0.2284	N/A	N/A
b[2,6,1]	0.1257	0.7366	0.03112
b[2,6,2]	0.8052	0.7578	0.03271
b[2,6,3]	-0.9309	N/A	N/A
b[2,7,1]	0.4726	0.7372	0.03113
b[2,7,2]	0.6983	0.7573	0.03263
b[2,7,3]	-1.1710	N/A	N/A
b[2,8,1]	0.9219	0.7358	0.03111
b[2,8,2]	0.2186	0.7585	0.03267
b[2,8,3]	-1.1400	N/A	N/A

Note. The number of MC draws used to compute these statistics is 100,000.
[a]b[t,j,k] represents the parameter estimate at the time point *t* (t = 1 for the pre-test while t =2 for the post-test) for the item *j* with the response category *k*.
[b]Since the parameters with response category 3 are not estimated here – they were obtained by computing the sum of parameter estimates with the first two response categories and then reversing the sign of the sum, the associated SDs and MC errors are not available.

Table 26.
The Model Elicited Given the Features of the Item before and after Instruction for the FMCE Data with 8 items

	1 (2)[a]	2 (5)	3(11)	4(12)	5(8)	6(9)	7(10)	8(13)
Before	2[b]	1	2	2	3	2	2	2
After	2	1	1	1 or 2	3	2	2 or 1	1

[a]The number in the parenthesis refers to the original item number.
[b]indicates the model elicited by the item.

We can compare the parameter estimates for the common items (the first 4 on FMCE8) between FMCE4 and FMCE8. Since the data were obtained from the same group of students, we expect that the parameter estimates for those items would be similar, with some variations due to sampling error or due to the fact that with FMCE8 more information is available about each examinee, and his/her propensity to use model can be better estimated. (Note that in our analysis we will not be able to evaluate which set of parameter estimates is better than the other since only one set of data is used and "true" values are not known. In general, this kind of evaluation can be done by a simulation study, but this lies beyond the concerns of the current study.)

By comparing Tables 22 and 25 in terms of the parameter estimates, we can see that differences exist. However, if we take into account the posterior standard deviations (labeled as SD in the table) associated with each parameter estimate, these discrepancies may not be statistically significant. The following discussion should thus be considered as descriptive rather than inferential. Except for item 4 (question 12), the relative position of parameter estimates within each item (i.e., the model the item tends to elicit) is the same between FMCE4 and FMCE8, as seen by comparing Table 23 with Table 26. Item 4, after instruction, tends to elicit both models 1 and 2 (b[2,4,1] and b[2,4,2] are 0.4393 and 0.4625, respectively) with FMCE4; with FMCE8, it has a greater tendency to provoke a response based on model 1, followed by model 2 (b[2,4,1] and b[2,4,2] are 0.6058 and 0.4310, respectively). Notice that for latter the difference between b[2,4,1] and b[2,4,2] is small, and in terms of practical concern they could be interpreted as effectively the same. That is, students are in the transition toward learning a new physics concept and have a

propensity to provide a response consistent with the use of either model 1 or 2, as with FMCE4 regarding item 4.

There are four additional items in FMCE8. Items 5 through 7 (corresponding to questions numbered 8 through 10 in FMCE) are in the same set of questions involving a toy car that is given a quick push so that it rolls up an inclined ramp. This scenario is in fact expert-equivalent to the one for items 3, 4, and 8 (corresponding to questions numbered 11, 12, and 13 in FMCE), the tossed coin as discussed earlier. Furthermore, items 5-7 ask the same kind of questions as items 3, 4, and 8 – "the car is moving up the ramp after it is released" (vs. "the coin is moving upward after it is released"), "the car is at its highest point" (vs. "the coin is at its highest point"); and "the car is moving down the ramp" (vs. "the coin is moving downward"), respectively. However, students do not respond consistently on those items, especially after instruction – no wonder, since as discussed above, students are in a mixed model state after instruction, and their responses are still influenced by features that are not relevant from the expert perspective.

From Table 26, we can see that before instruction, the difference between these two sets of questions occurs only for item 5. This item has a greater tendency to elicit a response based on model 3 (in contrast, item 3, its expert-equivalent question, tends to evoke a response based on model 2). This finding again suggests context dependence. After instruction, students retain their propensity to use null or incorrect models for items 5 through 7 (with an exception for item 7 since it also tends to evoke responses based on model 1), indicating that students still have difficulties in understanding force-motion concept or applying it to the situations depicted in items 5 through 7. The latter explanation is more likely, since its expert-equivalent set of questions (i.e., items 3, 4, and

8) tends to provoke responses mainly based on model 1, indicating that students at least learned the concept to the extent that they knew how to apply it to a tossed coin. The force involving a tossed coin is relatively straightforward – an up and down force, even if students do not know it is gravity. But the initial force applied to a toy car to make it move on the inclined ramp is a force with an angle, and therefore it involves the concept of "net force". The initial force can be decomposed into the horizontal and vertical force plus a gravitational force in the vertical direction. Students who do not know Newton's first law – an object can continue to move without force – think that they need to deal with the net force to make a correct answer, so they could easily to choose the answer, for example, stating that "the net force is zero" when the car is at its highest point, or the answer describing that "net increasing force down ramp" when the car is moving down the ramp. These responses are based on model 2, an incorrect model. (Again, the reasoning students used to respond items 5 through 7 can be further examined by interviewing them.) Because of students' inconsistency in responding to equivalent sets of questions, their propensity to use a certain model for problem-solving is in a mixed state with regard to the set of tasks encompassed by the FMCE8 data.

Acc5. The last data set deals with the concept of acceleration. Unlike the first three data sets, the partially homogeneous model is preferred (the *DIC* values for the homogeneous, partially homogeneous, and heterogeneous model are 2583.370, 2393.120, and 2402.910, respectively), indicating that although the population distribution has been changed after instruction (see Table 27) but one set of item parameter estimates before and after instruction would be adequate to describe the data. This could be due to the fact that students change their model use consistently. This reflects on items as well, as seen in

Table 28: before instruction every item tends to elicit a response based on model 2, while after instruction each has a greater tendency to evoke a response based on model 1. Therefore, instead of using two sets of item parameter estimates (one for pre-test and another for post-test) the partially homogeneous model with one set of item parameter estimates is preferred. Our discussion below regarding the items is therefore based on the partially homogeneous model.

Table 27.
The Average Parameter Estimates over Persons before and after Instruction for the Acceleration Data

Parameters	Mean	SD	MC error
(Pre-test)			
mu11[a]	-0.4382	0.3903	0.01401
mu12	1.1310	0.3798	0.01373
mu13	-0.6928	N/A[b]	N/A
(Post-test)			
mu21	1.6670	0.3902	0.01399
mu22	0.0534	0.3842	0.01373
mu23	-1.7204	N/A	N/A

Note. The number of MC draws used to compute these statistics is 100,000.
[a]mu*tk* represents the parameter estimate at the time point *t* (t = 1 for the pre-test while t =2 for the post-test) over persons with response category *k*.
[b]Since the parameters with response category 3 are not estimated here – they were obtained by computing the sum of parameter estimates with the first two response categories and then reversing the sign of the sum, the associated SDs and MC errors are not available.

Table 28.
The Item Parameter Estimates for the Acceleration Data under the Heterogeneous AR Model

Item Parameter	Mean	SD	MC error
(Pre-test)			
b[1,1,1][a]	0.5759	0.7397	0.02941
b[1,1,2]	1.2100	0.7297	0.02997
b[1,1,3]	-1.7860	N/A[b]	N/A
b[1,2,1]	-0.5645	0.7361	0.02943
b[1,2,2]	0.8261	0.7231	0.03001
b[1,2,3]	-0.2616	N/A	N/A
b[1,3,1]	-0.1900	0.7333	0.02944
b[1,3,2]	0.4300	0.7225	0.03000
b[1,3,3]	-0.2400	N/A	N/A
b[1,4,1]	-0.6352	0.7360	0.02945
b[1,4,2]	0.4897	0.7227	0.02999
b[1,4,3]	0.1455	N/A	N/A
b[1,5,1]	0.0234	0.7339	0.02945
b[1,5,2]	0.6604	0.7233	0.02997
b[1,5,3]	-0.6838	N/A	N/A
(Post-test)			
b[2,1,1]	1.5390	0.7606	0.02987
b[2,1,2]	0.6620	0.8026	0.03152
b[2,1,3]	-2.2010	N/A	N/A
b[2,2,1]	0.6862	0.7487	0.03000
b[2,2,2]	0.2362	0.7914	0.03171
b[2,2,3]	-0.9224	N/A	N/A
b[2,3,1]	1.0660	0.7502	0.03001
b[2,3,2]	-0.0926	0.7944	0.03167
b[2,3,3]	-0.9738	N/A	N/A
b[2,4,1]	0.6920	0.7483	0.03001
b[2,4,2]	0.1069	0.7925	0.03171
b[2,4,3]	-0.7988	N/A	N/A
b[2,5,1]	1.4110	0.7557	0.03001
b[2,5,2]	0.2837	0.7996	0.03167
b[2,5,3]	-1.6940	N/A	N/A

Note. The number of MC draws used to compute these statistics is 100,000.
[a]b[t,j,k] represents the parameter estimate at the time point *t* (t = 1 for the pre-test while t =2 for the post-test) for the item *j* with the response category *k*.
[b]Since the parameters with response category 3 are not estimated here – they were obtained by computing the sum of parameter estimates with the first two response categories and then reversing the sign of the sum, the associated SDs and MC errors are not available.

We summarize the item parameter estimates as seen in Table 29. Note that items 1 through 5 refer to questions 22 through 26 in FMCE – a set of questions used to measure students' understanding about the concept of acceleration. They involve a toy car which can move to the right or left on a horizontal surface along a straight line (the + distance axis). Each item describes the car with a different motion (e.g., item 1 states that the car moves toward the right, speeding up at a steady rate) and asks which choice representing the acceleration-time graph corresponds to the motion of car (see Appendix A for details).

Table 29.
The Item Parameter Estimates for the Acceleration Data under the Partially Homogeneous AR Model

Item Parameter	Mean	SD	MC error
b[1,1][a]	0.8475	0.3971	0.01379
b[1,2]	0.7295	0.3887	0.01360
b[1,3]	-1.5770	N/A[b]	N/A
b[2,1]	-0.1641	0.3902	0.01387
b[2,2]	0.3098	0.3809	0.01369
b[2,3]	-0.1457	N/A	N/A
b[3,1]	0.2067	0.3902	0.01391
b[3,2]	-0.0594	0.3808	0.01366
b[3,3]	-0.1473	N/A	N/A
b[4,1]	-0.2109	0.3900	0.01390
b[4,2]	0.0417	0.3805	0.01368
b[4,3]	0.1692	N/A	N/A
b[5,1]	0.4630	0.3912	0.01384
b[5,2]	0.2037	0.3816	0.01364
b[5,3]	-0.6667	N/A	N/A

Note. The number of MC draws used to compute these statistics is 100,000.
[a]b[j,k] represents the parameter estimate for the item *j* with the response category *k*.
[b]Since the parameters with response category 3 are not estimated here – they were obtained by computing the sum of parameter estimates with the first two response categories and then reversing the sign of the sum, the associated SDs and MC errors are not available.

From Table 29, we can see that item 1 tends to provoke a response consistent with model 1 or 2 (b[1,1] and b[1,2] are 0.8475 and 0.7295, respectively). It is not difficult to understand why this occurs. Students who use the correct model – speeding up at a steady rate means the acceleration is constant – would choose a graph that corresponds to a constant acceleration (i.e., choices A or B). Note that the correct answer in fact is choice A (not B) since the car moves to right, indicating that the sign of acceleration is positive. Those who choose B may not take into account the direction of the car's moving but at least demonstrate their understanding about what "speeding up at a steady rate" means in terms of acceleration. Therefore, those who choose A or B would be considered to use the correct model. On the other hand, if students do not understand the concept of acceleration, they would choose a graph that represents an increasing acceleration since the car is "speeding up" – they confuse the concepts of velocity and acceleration. Thus, this item has a greater tendency to evoke not only model 1 responses but also model 2 responses.

Item 2 differs from item 1 in that the car is not speeding up at a steady rate but slowing down at a steady rate. It tends to elicit a response based on model 2, the incorrect model. For the similar reasons as for item 1, students who think of acceleration as velocity would choose the graph(s) that corresponds to decreasing acceleration over time since the car is "slowing down." It is not clear, however, why items 1 and 2 (which can be considered as the same type of question) have different tendencies to evoke a certain model. Maybe students' comprehension about "slowing down" is not as straightforward or natural as "speeding up" when it comes to the graphical representation.

Item 3 has a greater tendency to elicit a response based on the correct model. It states that the car moves toward the left at a constant velocity. Students may have a better understanding about what "constant velocity" refers to: zero acceleration. Therefore, they have a greater propensity to use model 1 thinking on this item. Similarly, item 5, which is identical to item 3 but has the car moving toward the right (not left), has a greater tendency to provoke a response based on model 1 although it also tends to evoke responses based on model 2 (b[5,1] and b[5,2] are 0.4630 and 0.2037, respectively).

Finally, item 4, which is similar to item 1 but now has the car moving toward the left, has a different story to tell. It has a greater tendency to evoke a response based on model 3, the null or unsystematic model. Comparing this result with that of item 1, it is possible that students are more likely to be confused when the car moves toward the left than right for the same kind of reason that they are more used to a "speeding up" than "slowing down" scenario in terms of graphical representation. For naïve students, this item would be hard for them to respond intuitively, and therefore they tend to use a model 3 approach.

Chapter V

Summary and Conclusions

The goal of the current study was to use a formal psychometric model (i.e., the Andersen-Rasch multivariate measurement model, AR; Andersen, 1973 & 1995) to study students' conceptual understanding in physics (in particular, Newtonian mechanics). The perspective is based on the "evidence-centered" design (ECD; Mislevy, Steinberg, & Almond, 2003) framework. The study builds on the Force Concept Inventory (FCI; Hestenes, Wells, & Swackhammer, 1992) and the Force Motion Concept Evaluation (FCME; Thornton & Sokoloff, 1998) task design and on previous analyses of the cognitive processes of physics problem-solving. It thus focuses on the measurement component of evidence model (EM) in the ECD stage called the Conceptual Assessment Framework (CAF). The use of the AR model for tasks designed to reveal students' conceptions/misconceptions in physics is consistent with a cognitive perspective of learning, namely that students' solve problems using approaches that can often be identified with conceptions or common misconceptions, and their propensity to use a certain approach (in this case Newtonian, "impetus theory", or Aristotelian) for problem-solving depends on the features of the item presented to him/her. To demonstrate this, four data sets (one from FCI and the others from FMCE, labeled FCI5, FMCE4, FMCE8, and Acc5, respectively) were used and analyzed with the AR model using a Markov Chain Monte Carlo (MCMC; Gelman, Carlin, Stern, & Rubin, 1995) estimation procedure, carried out with the *BUGS* computer program (**B**ayesian inference analysis **U**sing **G**ibbs **S**ampling; Spiegelhalter, Thomas, & Gilks, 1997). We summarize the results,

discuss the limitations of the current study, and consider some potential research questions for the future.

Summary and Conclusions to the Psychometric Analysis

The first data, FCI5, contains students' responses about force-motion relation. Based on values of the Bayesian model-fit index *DIC* (Deviance Information Criterion; Spiegelhalter, Best, Carlin, & van der Linde, 2002) the heterogeneous AR model is preferred to the homogeneous and partially homogeneous AR models, indicating that both student distribution and item parameter estimates have been changed after instruction. Before instruction, students have a greater propensity to use model 2 (incorrect "impetus theory" conceptions of force and motion). After instruction, they still tend to use model 2 although their tendency to use model 1 (correct Newtonian conception of force and motion) has been increased. One can view this as most students being in a mixed model state; they may use model 1 or model 2, with probabilities that depend on the particular features of the task they are solving. By further studying the change in terms of item parameter estimates, we indicated two possible common misconceptions in their incorrect answers. One is related to Newton's first law (students tend to believe that there is a force acting on the object to keep it moving) and another is related to frictional forces. Since those questions in FCI5 were also used in Bao and Redish's studies, we compared our results with their findings. The two studies yield similar results with a slight difference in terms of model use after instruction: Their analysis indicated that both models 1 and 2 are about equally likely to be used by students.

The second data we used is FMCE4. As with FCI5, questions in FMCE4 measure students' understanding about forces and force-motion concepts. The heterogeneous model is again preferred, based on values of the *DIC*. Before instruction, students tend to use model 2 (the same as for FCI5); after instruction, however, students have a greater tendency to use model 1 rather than model 2 (although the difference is not substantial), indicating that most students have improved their understanding about force-motion relation to some extent, but not fully. This is confirmed by examining the shift of item parameter estimates before and after instruction. The item level analysis indicates that students still tend to believe that there is always a force in the direction of an object to keep it moving, the same misconception revealed by FCI5. In addition, students seem to incorrectly believe that when the tossed coin is at its highest point, the magnitude of gravity is zero. Comparing these results with Bao's study using questions in FMCE4, Boa's analysis shows a slight difference after instruction. Bao indicated that students tend to use either model 1 or 2 (more in model 2 but not by much).

FMCE8, the third data we used, has 4 additional items to FMCE4. As for the first two data, the heterogeneous model is selected for its smallest *DIC* value. On the pre-test, students have a greater propensity to use model 2 (the same as for FCI5 and FMCE4). After instruction, students are equally likely to use either model 1 or 2, strongly indicating that students are in a mixed model state. We then compared the item parameter estimates for the common items between FMCE4 and FMCE8. The analysis shows the similar results one would expect, after taking into account the posterior standard deviation associated with the parameters being estimated. Further examining the 4 additional items

implies that students, again, seem to believe that a force is needed to keep an object moving.

Based on what was found from the first three data sets (questions in those data are used to measurement students' understanding about force-motion relation in some ways although the item format is different between FCI and FMCE – i.e., the former consists of five choices for each item, while for latter it has seven or eight choices associated with each item), as well as comparing with what Bao and Redish found in their studies, it appears that these students had a greater propensity to give responses consistent with model 2 before instruction, and they give responses consistent with either model 1 or 2 (more or less, depending on the features of the item and characterized by the item parameters in the AR model) after instruction. This indicates students are in a mixed model state (i.e., in a transition toward understanding Newtonian mechanics) after one semester of physics learning. This also implies that they still have difficulties in understanding some concepts related to force-motion. In particular, they incorrectly believe that there is a force acting on an object to keep it continue to move, one of common misconceptions identified by Bao and Redish (2004). However, the particular features of tasks still evoke different response categories to expert-equivalent items, indicating that the students are still not in a "pure state" of Newtonian responding.

Finally, we analyzed the acceleration data, Acc5. Unlike the first three data sets, the partially homogeneous model is preferred, indicating that although the population distribution has changed, a single set of item parameter estimates appears to be adequate to describe the data. Before instruction, students tend to provide responses consistent with model 2 for problem-solving. After instruction, however, they have a greater propensity to

provide responses consistent with model 1, implying that students have improved their understanding about the concept of acceleration. We further examined the item parameter estimates for each item and found that students may have difficulties to represent their understanding about acceleration in terms of acceleration-time graphs, especially when the object is slowing down or moving toward the left, in which case the sign of acceleration in both task scenarios is negative.

It can be seen the analysis based on the psychometric AR model provides additional information beyond the Bao and Redish analyses (Bao, 1999; Bao & Redish, 2001 & 2004) both at the item and person level. First, the vector-valued parameter estimates for each item shows its tendency to provoke the response based on a certain model (or models); similarly, the vector-valued parameter estimates for each person indicates his/her propensity to use a specific model (or models) given the features of the items. Second, the accompanying standard deviations and MC errors can be used to estimate the accuracy of parameter estimates. Third, the shift of parameter estimates (in terms of item or persons) from before to after instruction implies what kind of learning has taken place and what concepts students seem to have difficulties in understanding, which can help to improve the physics instruction. Fourth, carrying out analysis within the formal framework of probability-based reasoning allows us to use criteria (e.g., *DIC* used in the current study) to assess the fit of the model to the data as well as to compare different models (e.g., the homogeneous, partially homogeneous, and heterogeneous AR models compared in this study).

Limitations of the Current Study

When making inferences from this study, one should be aware that the sample size for each data set is relatively small, and most of students are biology majors. They may not be a good representation of other student populations, and may differ from identifiable groups such as physics and engineering majors or humanities majors. Therefore, the results of this study need to be interpreted with caution. Because of the relatively small sample size (by the standards of psychometric analyses such as latent class and IRT modeling), not much information is available to estimate parameters for individual items and students. Particularly with the partially homogeneous and heterogeneous models, model parameters were less precisely estimated, and it usually took many more MCMC cycles to converge.

Directions for Future Research

As discussed in Chapter II, there are four perspectives to study the nature of human mind: the differential, behaviorist, cognitive, and situative perspectives. In the past, most research has focused on the first two perspectives, while the last two approaches have not been explored in great detail, especially in the field of educational measurement. The current study explores students' learning in terms of cognitive perspective. However, we only focus on a small piece of cognitive process (i.e., the mixture-within-persons strategy for problem-solving). For this aspect of learning, modeling students' problem-solving in terms of common misconceptions has proven useful. Other fields of learning may not be the same as physics in this regard, so they may focus on different aspects of cognitive process. In this study, we have provided a good example of integrating ideas from the

cognitive science of learning and modeling students' responses in the ECD framework to examine students' problem-solving in physics. We can continue to follow this line of research to explore, for example, how students solve a class of math tasks and how it is different from physics learning.

As also discussed in Chapter II, most current testing practices are not a good match with the situative perspective. More studies are needed to explore how social or cultural factors affect students' learning. For example, one may be interested in investigating whether the different test formats (multiple-choice test vs. open-ended questions) affect students' problem-solving strategy and how it occurs. Again, in principle this research can be carried out in the ECD framework. It is likely that extensions of interpretations and of psychometric models themselves will be called for, just as the AR extends beyond the overall-proficiency models that are used for most testing applications.

Finally, one might consider extending the AR model by incorporating, for example, the category weight to further examine its use in terms of studying mental model states. This extension of the AR model could take the following form:

$$P(X_{ij} = p) = \exp(\omega_p \theta_{ip} + \beta_{jp}) / \sum_{p=1}^{m} \exp(\omega_p \theta_{ip} + \beta_{jp}), \qquad (28)$$

where:

ω_p is the category weight associated with the response category p.

Such a model would provide for more variety in the shapes of conditional probability distributions, but questions of identifiability and estimability would need to be explored.

Appendix A

Physics Questions

Questions for the FCI data – the 1st data set:

5. A boy throws a steel ball straight up. Discarding any effects of air resistance, the force(s) acting on the ball until it returns to the ground is (are):

   a) its weight vertically downward along with a steadily decreasing upward force.
   b) a steadily decreasing upward force from the moment it leaves the hand until it reaches its highest point beyond which there is a steadily increasing downward force of gravity as the object gets closer to the earth.
   c) a constant downward force of gravity along with an upward force that steadily decreases until the ball reaches its highest point, after which there is only the constant downward force of gravity.
   d) a constant downward force of gravity only.
   e) none of the above, the ball falls back down to the earth simply because that is its natural action.

9. The main forces acting, after the "kick", on the puck along the path you have chosen are:

   a) the downward force due to gravity and the effect of air pressure.
   b) the downward force of gravity and the horizontal force of momentum in the direction of motion.
   c) the downward force of gravity, the upward force exerted by the table, and a horizontal force acting on the puck in the direction of motion.
   d) the downward force of gravity and an upward force exerted on the puck by the table.
   e) gravity does not exert a force on the puck, it falls because of intrinsic tendency of the object to fall to its natural place.

18. An elevator, as illustrated (skipped), is being lifted up an elevator shaft by a steel cable. When the elevator is moving up the shaft at a constant velocity:

   a) the upward force on the elevator by the cable is greater than the downward force of gravity.
   b) the amount of upward force on the elevator by the cables equals that of the downward force of gravity.
   c) the upward force on the elevator by the cable is less than the downward force of gravity.
   d) it goes up because the cable is being shortened, not because of the force being exerted on the elevator by the cable.
   e) the upward force on the elevator by the cable is greater than the downward force due to the combined effects of air pressure and the force of gravity.

22. A golf ball driven down a fairway is observed to travel through the air with a trajectory (flight path) similar to that in the depiction below (skipped). Which following force(s) is(are) acting on the golf ball during its entire flight?

1. the force of gravity
2. the force of the "hit"
3. the force of air resistance

a) 1 only
b) 1 and 2
c) 1, 2, and 3
d) 1 and 3
e) 2 and 3

28. A large box is being pushed across the floor at a <u>constant speed</u> of 4.0 m/s. What can you conclude about the forces acting on the box?

a) If the force applied to the box is doubled, the constant speed of the box will increase to 8.0 m/s.
b) The amount of force applied to move the box at a constant speed must be more than its weight.
c) The amount of force applied to move the box at a constant speed must be equal to the amount of the frictional forces that resist its motion.
d) The amount of force applied to move the box at a constant speed must be more than the amount of the frictional force that resist its motion,
e) There is a force being applied to the box to make it move but the external forces such as friction are not "real" forces they just resist motion.

<u>Questions for the Force-Motion data with 4 items – the 2nd data set</u>:

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 <u>one</u> 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 no at all but choose only one answer for each blank. If you think that none is correct, answer choice **J**.

(graphs are skipped here)
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).

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

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?

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 case describe 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. (This item is used for the 3rd data set only)

Questions for the Force-Motion data with 8 items – the 3rd data set
(note: only those items addition to the 2nd data set are listed)

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 that it can be ignored*. (the graph is skipped here)

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 for the Acceleration data – the 4th data set

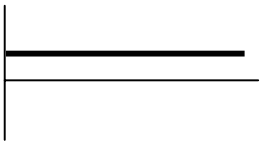
Questions 22-26 refer to a toy car which can move to the right or left on a horizontal surface along a straight line (the + distance axis). The positive direction is to the right.

(The graph is skipped here)

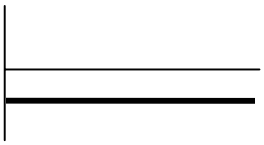
Different motions of the car are described below. Choose the letter (**A** to **G**) of the acceleration-time graph which corresponds to the motion of the car described in each statement. (note: the horizontal axis represents time while the vertical axis represents the acceleration; and the interception between the horizontal and vertical axis is zero.)

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

A.

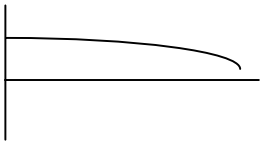


B.

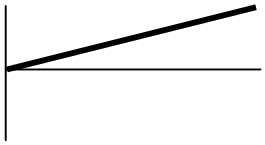


C.

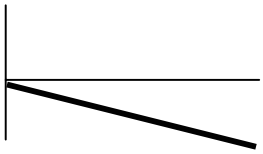
D.



E.



F.



G.

J. None is correct.

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.

Appendix B

Associations between the Physics Models and the Choices

For the 1st data set – 5 FCI questions on Force-Motion concept

Questions	Model 1	Model 2	Model 3
5	d	a, b, c	e
9	a, d	b, c	e
18	b	a, e	c, d
22	a, d	b, c, e	N/A
28	c	a, d, e	b

For the 2nd data set – 4 FMCE questions on Force-Motion concept

Questions	Model 1	Model 2	Model 3
2	d	b	others
5	d	b	others
11	a	g	others
12	a	d	others

For the 3rd data set – 8 FMCE questions on Force-Motion concept

Questions	Model 1	Model 2	Model 3
2	d	b	others
5	d	b	others
11	a	g	others
12	a	d	others
8	a	g	others
9	a	d	others
10	a	b	others
13	a	b	others

For the 4th data set – 5 FMCE questions on Acceleration concept

Questions	Model 1	Model 2	Model 3
22	a, b	e, f	others
23	a, b	f, g	others
24	c	a, b	others
25	a, b	e, f	others
26	c	a, b	others

Appendix C

The BUGS Codes for Estimating Parameters under the Homogeneous, Partially Homogeneous, and Heterogeneous AR models (Using the 1st Data Set only)

The BUGS code for the homogeneous model

```
# Andersen's multivariate Rasch model:
# the first 99 responses are from the pre-test and the remaining 98 are from the post-test
#
# scale fixed by centering parameters for each item around zero &
# also by centering parameters for each person around zero
#
# Model A -- One BUGS run, with the same conditional probabilities and the same
# examinee population distributions for thetas over all subjects & time # points.
```

Model

```
{
   for (j in 1:ni){
    b[j,1] ~ dnorm(0,1)
    b[j,2] ~ dnorm(0,1)
    b[j,3] <- -(b[j,1]+b[j,2])
   }

   for (i in 1:N){
    for (j in 1:ni){
      for (k in 1:3){
        x[i,j,k] <- exp(theta[i,k] + b[j,k])
      }
      sum[i,j] <- sum(x[i,j,1:3])
      for (l in 1:3){
        p[i,j,l] <- x[i,j,l]/sum[i,j]
      }
      resp[i,j] ~ dcat(p[i,j,1:3])
    }
    theta[i,1] ~ dnorm(0,1)
    theta[i,2] ~ dnorm(0,1)
    theta[i,3] <- -(theta[i,1]+theta[i,2])
   }
}
```

```
#inits
list(b = structure(.Data=c(
.1,.9,NA,
.6,.5,NA,
.1,.05,NA,
2.0,2.0,NA,
.5,2.5,NA), .Dim=c(5,3)))

list(b = structure(.Data=c(
.8,.2,NA,
.2,.8,NA,
.5,.5,NA,
.2,.2,NA,
1.5,.5,NA), .Dim=c(5,3)))

#data
list(N = 197,
ni = 5,
resp = structure(.Data = c(
2,3,3,1,2,
1,3,3,1,2,
2,3,3,2,2,
2,1,1,1,2,
1,2,3,NA,NA,
.
.
.

1,1,3,1,2,
2,3,1,2,2,
2,3,1,2,2,
2,3,2,1,3,
2,1,3,2,2), .Dim = c(197, 5)))
```

The BUGS code for the partially homogeneous model

```
# Andersen's multivariate Rasch model:
# the first 99 responses are from the pre-test and the remaining 98 are from the post-test
#
# scale fixed by centering parameters for each item around zero &
# also by centering parameters for each person around zero
#
# Model B -- One BUGS run, with the same conditional probabilities for all
# subjects and time points but different population distributions for
# pre-test response data and post-test response data


Model

{
   for (j in 1:ni){
     b[j,1] ~ dnorm(0,1)
     b[j,2] ~ dnorm(0,1)
     b[j,3] <- -(b[j,1]+b[j,2])
   }

# for pre-test data (t=1)

   for (t in 1:1){
     for (i in 1:99){
      for (j in 1:ni){
        for (k in 1:3){
          x[i,j,k] <- exp(theta[t,i,k] + b[j,k])
        }
        sum[i,j] <- sum(x[i,j,1:3])
        for (l in 1:3){
          p[i,j,l] <- x[i,j,l]/sum[i,j]
        }
        resp[i,j] ~ dcat(p[i,j,1:3])
      }
      theta[t,i,1] ~ dnorm(mu11,1)
      theta[t,i,2] ~ dnorm(mu12,1)
      theta[t,i,3] <- -(theta[t,i,1]+theta[t,i,2])
     }
     mu11 ~ dnorm(0,1)
     mu12 ~ dnorm(0,1)
   }
```

```
# for post-test data (t=2)

   for (t in 2:2){
     for (i in 100:N){
      for (j in 1:ni){
        for (k in 1:3){
          x[i,j,k] <- exp(theta[t,i,k] + b[j,k])
        }
        sum[i,j] <- sum(x[i,j,1:3])
        for (l in 1:3){
          p[i,j,l] <- x[i,j,l]/sum[i,j]
        }
        resp[i,j] ~ dcat(p[i,j,1:3])
      }
      theta[t,i,1] ~ dnorm(mu21,1)
      theta[t,i,2] ~ dnorm(mu22,1)
      theta[t,i,3] <- -(theta[t,i,1]+theta[t,i,2])
     }
     mu21 ~ dnorm(0,1)
     mu22 ~ dnorm(0,1)
   }
}

#inits
list(b = structure(.Data=c(
.1,.9,NA,
.6,.5,NA,
.1,.05,NA,
2.0,2.0,NA,
.5,2.5,NA), .Dim=c(5,3)))

list(b = structure(.Data=c(
.8,.2,NA,
.2,.8,NA,
.5,.5,NA,
.2,.2,NA,
1.5,.5,NA), .Dim=c(5,3)))

#data
list(N = 197,
ni = 5,
resp = structure(.Data = c(
2,3,3,1,2,
1,3,3,1,2,
2,3,3,2,2,
2,1,1,1,2,
```

1,2,3,NA,NA,
.
.
.

1,1,3,1,2,
2,3,1,2,2,
2,3,1,2,2,
2,3,2,1,3,
2,1,3,2,2), .Dim = c(197, 5)))

The BUGS code for the heterogeneous model

```
# Andersen's multivariate Rasch model:
# the first 99 responses are from the pre-test and the remaining 98 are from the post-test
#
# scale fixed by centering parameters for each item around zero &
# also by centering parameters for each person around zero
#
# Model C -- One BUGS run, different item parameters and population
# distributions for pre-test response data and post-test response data


Model

{

# for pre-test data (t=1)

   for (t in 1:1){
    for (j in 1:ni){
      b[t,j,1] ~ dnorm(mub11,1)
      b[t,j,2] ~ dnorm(mub12,1)
      b[t,j,3] <- -(b[t,j,1]+b[t,j,2])
    }
    mub11 ~ dnorm(0,1)
    mub12 ~ dnorm(0,1)

    for (i in 1:99){
      for (j in 1:ni){
        for (k in 1:3){
          x[i,j,k] <- exp(theta[t,i,k] + b[t,j,k])
        }
        sum[i,j] <- sum(x[i,j,1:3])
        for (l in 1:3){
          p[i,j,l] <- x[i,j,l]/sum[i,j]
        }
        resp[i,j] ~ dcat(p[i,j,1:3])
      }
    theta[t,i,1] ~ dnorm(mu11,1)
    theta[t,i,2] ~ dnorm(mu12,1)
    theta[t,i,3] <- -(theta[t,i,1]+theta[t,i,2])
    }
    mu11 ~ dnorm(0,1)
    mu12 ~ dnorm(0,1)
   }
```

```
# for post-test data (t=2)

  for (t in 2:timept){
    for (j in 1:ni){
      b[t,j,1] ~ dnorm(mub21,1)
      b[t,j,2] ~ dnorm(mub22,1)
      b[t,j,3] <- -(b[t,j,1]+b[t,j,2])
    }
    mub21 ~ dnorm(0,1)
    mub22 ~ dnorm(0,1)

    for (i in 100:N){
      for (j in 1:ni){
        for (k in 1:3){
          x[i,j,k] <- exp(theta[t,i,k] + b[t,j,k])
        }
        sum[i,j] <- sum(x[i,j,1:3])
        for (l in 1:3){
          p[i,j,l] <- x[i,j,l]/sum[i,j]
        }
        resp[i,j] ~ dcat(p[i,j,1:3])
      }
    theta[t,i,1] ~ dnorm(mu21,1)
    theta[t,i,2] ~ dnorm(mu22,1)
    theta[t,i,3] <- -(theta[t,i,1]+theta[t,i,2])
    }
    mu21 ~ dnorm(0,1)
    mu22 ~ dnorm(0,1)
  }

}

#inits
list(b = structure(.Data=c(
.2,.8,NA,
.7,.4,NA,
.2,.2,NA,
2.5,2.5,NA,
.6,3,NA,
.1,.9,NA,
.6,.5,NA,
.1,.05,NA,
2.0,2.0,NA,
.5,2.5,NA), .Dim=c(2,5,3)))
```

```
list(b = structure(.Data=c(
.8,.2,NA,
.4,.7,NA,
.2,.2,NA,
2.5,2.5,NA,
3,.6,NA,
.9,.1,NA,
.5,.6,NA,
.05,.1,NA,
2.0,2.0,NA,
2.5,.5,NA), .Dim=c(2,5,3)))

#data
list(N = 197,
timept = 2,
ni = 5,
resp = structure(.Data = c(
2,3,3,1,2,
1,3,3,1,2,
2,3,3,2,2,
2,1,1,1,2,
1,2,3,NA,NA,
.
.
.

1,1,3,1,2,
2,3,1,2,2,
2,3,1,2,2,
2,3,2,1,3,
2,1,3,2,2), .Dim = c(197, 5)))
```

REFERENCES

Adams, R. J., Wilson, M., & Wang, W. –C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21(1)*, 1-23.

Albert, J.H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17,* 251-269.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki, (Eds.), 2nd International Symposium on information Theory, Budapest: Akademiai Kiádo, pp. 267-281. Reprinted in S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Vol. 1. Foundations and basic theory*. New York: Springer-Verlag.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317-332.

Andersen, E.B. (1973). *Conditional inference and models for measuring.* Copenhagen: Danish Institute for Mental Health.

Andersen, E. B. (1995). Polytomous Rasch models and their estimation. In G. H. Fischer & I. W. Molen (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 271-291). New York: Springer-Verlag.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement, 22(2)*, 153-169.

Bao, L. (1999). *Dynamics of student modeling: A theory, algorithms, and applications to quantum mechanics*. Unpublished doctoral dissertation, University of Maryland, College Park, MD.

Bao, L. Hogg, K, & Zollman, D. (2002). Model analysis of fine structure of student models: An example with Newton's third law, accepted for publication by *Physics Education Research Section of the AJP*.

Bao, L., & Redish, E. F. (2001). Concentration analysis: A quantitative assessment of student states. *Physics Education Research Section of American Journal of Physics, 69 (7)*, 45-53.

Bao, L. & Redish, E. F. (2004). Educational assessment and underlying models of cognition. In W. E. Becker & M. Andrews (Eds), *The scholarship of teaching and learning in higher education: Contributions of research universities* (chap. 11). Indiana University Press.

Beguin, A. A. & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66(4)*, 541-561.

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

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference* (2nd ed.). New York: Springer.

Champagne, A., Klopfer, L., & Anderson, J. H. (1980). Factors influencing the learning of classical mechanics. *American Journal of Physics, 48(12)*, 1074-1079.

Dayton, C. M. (1998). *Latent class scaling analysis* (Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-126). Thousand Oaks, CA: Sage.

diSessa, A. (1982). Unlearning Aristotelian physics: A study of knowledge-based learning. *Cognitive Science, 5,* 37-75.

diSessa, A. (1993). Towards an epistemology of physics. *Cognition and Instruction, 10*, 105-225.

Driver, R. (1973). *The representation of conceptual frameworks in young adolescent science students*. Unpublished doctoral dissertation, University of Illinois, Urbana, Illinois.

Embretson, S.E. (Ed.) (1985). *Test design: Developments in psychology and psychometrics.* Orlando: Academic Press.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495-515.

Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement, 32(3)*, 277-294.

Embretson, S.E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Fischer, G. H. (1973). The linear logistic model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molen (Eds.), *Rasch models: foundations, recent developments, and applications* (pp. 131-155). New York: Springer-Verlag.

Fox, J –P. & Glas, C. A. W. (2001). Bayesian estimations of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66(2)*, 271-288.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 131-144). London: Chapman and Hall.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in practice*. London: Chapman & Hall.

Glas, C. A. W., & Falcón, J. C. S. (2003). A comparison of item-fit statistics. *Applied Psychological Measurement, 27(2)*, 87-106.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molen (Eds.), *Rasch models: foundations, recent developments, and applications* (pp. 69-96). New York: Springer-Verlag.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner and R. C. Calfee (Eds), *Handbook of educational psychology* (pp. 15-46). New York: Macmillan.

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

Hestenes, D., Wells, M., & Swackhammer, G. (1992). Force Concept Inventory. *The Physics Teacher, 30*, 141-153.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62*, 171-189.

Kim, S –H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement, 25(2)*, 163-176.

Langeheine, R. & van de Pol, F. (1994). State mastery learning: Dynamic models for longitudinal data. *Applied Psychological Measurement, 18(3)*, 277-291.

Lawson, R. A. (1984). *Student understanding of single particle dynamics*. Unpublished doctoral dissertation, University of Washington, Seattle, WA.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 4*, 149-174.

MathSoft, Inc. (1995). S-PLUS (Version 3.3 for Windows) [Computer program]. Seattle WA: Author.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science 210(4474)*, 1139-1141.

Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply imputed data sets. *Biometrika, 79*, 103-112.

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

Minstrell, J. (1982). Explaining the "at rest" condition of an object. *The Physics Teacher*, *20(1)*, 10-14.

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

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville IN: Scientific Software.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55(2)*, 195-215.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement, 1,* 3-67.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2)*, 159-176.

National Research Council (1999). *How people learn: Brain, mind, experience, and school.* Committee on Developments in the Science of Learning. Bransford, J. D., Brown, A. L., and Cocking, R. R. (Eds.). Washington, DC: National Academy Press.

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

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24(2)*, 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24(4)*, 342-366.

Rasch, G. (1961). On general laws and meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability* (Vol. 4). Berkeley: University of California Press.

Reiner, M., Slotta, J. D., Chi, M. T. H.,& Resnick, L. B. (2000). Naive physics reasoning: A commitment to substance-based conceptions. *Cognition and instruction, 18(1)*, 1-34.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology, 44*, 75-92.

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika, 60(4)*, 549-572.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.

Sjøberg, S., & Lie, S. (1981). Ideas about force and movement among Norwegian pupils and students. *Technical Report*, 81-11, Institute of Physics Report Series, University of Oslo.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of Royal Statistics, Soc. B, 64*, 583-640.

Steinberg, R., & Sabella, M. (1997). Performance on multiple-choice diagnostics and complementary exam problems. *Physics Teacher, 35*, 150-155.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51(4)*, 589-601.

Spiegelhalter, D.J., Thomas, A., Best, N.G., & Gilks, W.R. (1997). BUGS: Bayesian inference using Gibbs sampling (Version 0.60) [Computer program]. Cambridge, UK: University of Cambridge, Institute of Public Health, Medical Research Council Biostatistics Unit.

The MathWorks, Inc. (1996). MATLAB: The language of technical computing [Computer program]. Natick MA: Author.

Thornton, P. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation. *American Journal of Physics, 66(4)*, 338-351.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*, 371-390.

van der Heijden, P. G. M., Dessens, J., & Bockenholt, U. (1996). Estimating the concomitant-variable latent-class model with the EM algorithm. *Journal of Educational and Behavioral Statistics, 21(3)*, 215-229.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-139.

Vermunt, J. K., Langeheine, R., & Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics, 24(2)*, 179-207.

Viennot, L. (1979). Spontaneous reasoning in elementary dynamics. *European Journal of Science Education, 1(2)*, 205-222.

Volodin, N., & Adams, R. J. (1995). *Identifying and estimating a D-dimensional Rasch model*. Unpublished manuscript, Australian Council for Educational Research, Camberwell, Victoria, Australia.

Whitaker, R. (1983). Aristotle is not dead: Student understanding of trajectory motion. *American Journal of Physics, 51(4)*, 352-357.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement, 16(4)*, 309-325.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.