EXAMINING THE PSYCHOMETRIC FUNCTIONALITY OF THE FORCE CONCEPT INVENTORY

by

Philip Dale Eaton

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Physics

MONTANA STATE UNIVERSITY
Bozeman, Montana

April 2020

ACKNOWLEDGEMENTS

First, I would like to thank my mother and father, Kathryn and Dr. Timothy Eaton. They have always been encouraging me to pursue what I love and to never give up. Secondly, I have to thank my siblings, and their spouses, for looking out for me and for the completely unneeded enthusiasm whenever I talk about my work during family gatherings.

I thank my advisor Dr. Shannon Willoughby for allowing me to work with her on my Ph.D. thesis/dissertation. Shannon's laid-back attitude helped to keep me sane while work wanted to drive me crazy.

Further, I would like to thank all the individuals in the Montana State University Physics department who have allowed to me opportunities which have had profound impacts on my growth as a physicist, as an instructor, and as a person. Special thanks here to Hannah Gilroy, Margaret Jarrett, Sarah Barutha, and Dr. Yves Idzerda.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

LIST OF TABLES – CONTINUED

LIST OF FIGURES

LIST OF FIGURES – CONTINUED

Figure Page

ABSTRACT

To improve the current understanding of the Force Concept Inventory (FCI), both a response-option-level analysis and a dimensionality analysis were proposed and applied. The response-option-level analysis used polytomous item response theory to reveal that the response options on the FCI are generally functioning appropriately, with two questions being identified as likely malfunctioning. To address the question of the FCI's dimensionality, an analysis of local item independence using item response theory was proposed and performed. Results indicate that the FCI is a multi-factor instrument, not a unidimensional instrument as it is often assumed. As a result of this analysis, three factor models were proposed and tested using confirmatory factor analysis and confirmatory multi-trait item response theory. All of these models were found to adequately explain the factor structure of the FCI within each of the statistical frameworks. The results from these investigations can be used as a starting point for further analysis and directing future improvements of the FCI.

CHAPTER ONE

INTRODUCTION

The Physics Education Research (PER) field has made significant advances in the understanding of effective teaching pedagogies, specifically for instructing physics [5]. One of the most significant conclusions made by PER suggests that including active learning/interaction/engagement in the classroom at all levels of education often leads to statistically significant increases in the conceptual gains of students [5, 31]. Most of the evidence used to support this claim, and other similar claims are established on the results of preinstruction to postinstruction student performance on conceptual assessments. The most famous study to suggest the effectiveness of active learning was authored by Dr. Richard Hake in 1998 [31]. Within, Hake showed that comparing preinstruction to postinstruction student scores (through what is now called the Hake gain) revealed that active learning augmented pedagogy generally outperformed passive learning pedagogy (i.e., traditional lecturing) in a statistically significant manner for introductory physics classes. Hake's study served as a flashpoint in the PER field, after which many studies were performed to investigate all kinds of active learning strategies within all topics of physics. These investigations also extended into many different Science, Technology, Engineering, and Mathematics (STEM) fields, see Ref. [5] and the many references within for more details.

Like in Hake's study, the majority of evidence used to support conclusions made in PER comes from student responses to validated instruments. These instruments, also called assessments or evaluations, can take the form of a series of

free-response questions, multiple-choice questions, Likert-style questions, etc. Free-response questions are ones in which students have the ability to answer using pictures, sentences, or paragraphs, and are generally used to probe students' conceptual and mathematical skills without the limitation of having limited, specific response options available. Questions with limited response options, that students select one or more options from, are categorized as multiple-choice questions. Likert-style questions often contain 5 to 7 options for a scale ranging from "Strongly Agree" to "Strongly Disagree", and are generally used on self-report assessments to get a more holistic view of how students feel.

These types of questions are used to measure specific, targeted domains of knowledge. These domains include students' conceptual understanding of specific topics in physics and students' epistemological beliefs about physics. Typically, the questions that compose these instruments contain responses that can be scored or ranked in a statistically meaningful manner. Alternatively, qualitative questions could be used, but questions of this nature generally do not generate statistics used for statistical inference. Instead, qualitative questions are used to explore opinions and generally do not have a single "correct" answer. As a result, if a researcher intends to provide statistical evidence for any conclusions drawn within a study, then they will use an instrument that is comprised of questions with scored or ranked outputs.

There exist many instruments within PER whose questions output meaningful scores or ranks and probe a plethora of topics. The assessments currently used in PER can be loosely broken up into the specific domains of interest that they were designed to investigate, such as conceptual reasoning, mathematical skills, and epistemic beliefs. Some conceptual instruments currently used in PER include the Force Concept Inventory (FCI) [38], the Force and Motion Conceptual Evaluation (FMCE) [97], the Conceptual Survey of Electricity and Magnetism (CSEM) [60], the

Brief Electricity and Magnetism Evaluation (BEMA) [21], and the Thermal Concept Evaluation [105]. Assessments like these are used to gauge student understanding of specific concepts. Some epistemic instruments currently in use are the Colorado Learning Attitudes about Science Survey (CLASS) [1] and the Epistemological Beliefs Assessment for Physical Sciences (EBAPS) [27], which probe students' thoughts and beliefs about learning and thinking in science classes. There are many more instruments used in PER than the limited list supplied here, for more instruments see PhysPort **??**.

The difference between the instruments used in PER and exams written for a classroom to assess the knowledge of students in the context of course materials, is that PER instruments undergo a significant psychometric examination. Psychometrics is the science and statistics of understanding i) how an instrument functions on a particular population and ii) what student responses mean to researchers using an instrument. At first, these points appear to be entirely intertwined with one another. A different way to present these points to moderately disentangle them is: Psychometrics is the science and statistics of extracting i) information about the instrument *independent* of the students and ii) information about the students *independent* of the instrument. Some psychometric theories are incapable of entirely separating student information from test information, like Classical Test Theory and Factor Analysis [20, 45], and other theories are statistically formed to make student and instrument information independent, like Item Response Theory and Rasch Modeling [18].

Before an instrument is used in active research, it should undergo significant psychometric testing to ensure the functionality of the instrument is well understood. Armed with this information, researchers can properly select an instrument to use in their studies while being informed of how to use the instrument properly, what

the instrument is measuring, and any problems or biases inherent to the instrument. Attempting to interpret student responses to an instrument without this kind of understanding can lead to several issues, like i) incorrect conclusions being made about the student sample being analyzed, ii) inaccurate statistics being calculated, or iii) using wrong grading models to generate student scores. A case study of proper and improper psychometric practices can be seen in the history of the FCI.

The FCI was created to measure students' conceptual understanding of Newtonian Mechanics at the introductory level (i.e., Algebra- and Calculus-based Physics I). Specifically, the creators of the FCI were intending to measure students' conceptual understanding of Newton's first law, Newton's second law, Newton's third law, kinematics, and some vector properties. As an individual student's ability within each of these conceptual domains is not directly observable (i.e., one cannot *directly* measure a student's Newtonian ability) these disciplines are said to each represent different latent traits that the student has. Latent traits are not directly observable traits of subjects (i.e., students), such as intelligence, that are inferred through manifest observable variables (i.e., total scores on an assessment). From a psychometric point-of-view, this suggests that the FCI was created with a desired multiple latent trait structure (also referred to as a factor structure) designed to probe the previously indicated Newtonian concepts.

The first version of the FCI was written in 1992, however, it was not until 1995 that its factor structure was investigated using psychometric tools. In 1995, Huffman and Heller used exploratory factor analysis (EFA) to examine the statistical structure of the FCI [42]. Exploratory FA is a data-driven statistical method used to explore the underlying latent trait structure of a set of variables. This study found that the 1992 version of the FCI was only capable of probing 1-3 factors, depending on the dataset used. Meaning, the 1992 version of the FCI likely only probed 1-3 distinctly different

domains of knowledge, and not the 6 initially designed concepts [38]. In response to this study, the creators of the FCI suggested that including novice thinkers in an EFA would inadvertently alter the detected factor structure of the FCI [36]. They argued that the detected structure would be a combination of the true structure of the FCI and the structure of the FCI as interpreted by novice-like and expert-like thinkers within the datasets. Because the structure of the FCI was designed to reflect an expert categorization of Newtonian thinking, the inclusion of the novice-like thinkers would significantly impact the detected structure. Huffman and Heller responded to this complaint, using only the suggested expert-like thinkers in conjunction with EFA, and again found that the 1992 version of the FCI only probed 1-3 factors [34]. Incidentally, it turns out that EFA is not the correct tool to use for confirming proposed factor structures. The proper tool, classically, is Confirmatory Factor Analysis, which is discussed in Chapter 4 on page 89.

The original Huffman and Heller article and the subsequent response articles have come to be known as the "H&H Debate," and naturally called into the question the validity and functionality of the FCI, both the 1992 version and the updated 1995 version. As a result of this debate, researchers assumed the FCI's factor structure was dependent on the sample used to explore the instrument. This implies that the factor structure of the FCI is not stable, and thus does not measure the conceptual domains it originally was designed to probe. An argument can be made where this conclusion makes sense since all of the concepts on the FCI can be placed under the umbrella of Newtonian Mechanics. For example, novice-like students may master a single concept, like Newton's third law, while still not mastering the other concepts probed by the instrument. Within a classroom, there will exist many groups of students who have mastered some of the concepts probed, but not all. These groups could be uncovered using EFA. But, as students learn the material through instruction, their

thinking will become more expert-like. As a result, the number of factors detected using postinstruction data may appear to be less than the number of factors detected using preinstruction data. This is to say, as students become more expert-like in their thinking, the number of concepts extracted in EFA may decrease, tending towards a single factor (i.e., just Newtonian understanding). However, for an instrument to give reliable and reproducible results, its structure must be consistent across multiple student samples at all level of novice-like to expert-like thinking.

The structure of the updated 1995 FCI was not examined until 2012, about 17 years after it was released, when Scott, Schumayer, and Gray presented their EFA results [85]. They found a factor structure that suggested the 1995 FCI was probing 5 conceptual factors, which resembled the originally intended factor structure [38]. These results were further validated by Scott and Schumayer using exploratory multitrait item response theory (EMIRT) [82]. It should be noted that exploratory analyses only, as the name suggests, *explores* the factor structure of instruments to help propose factor models. That is, none of the aforementioned studies have tested proposed factor models for the FCI in a confirmatory manner.

From 1995 to 2012, the factor structure of the 1995 version of the FCI was not statistically explored. As a result, many studies published during this time investigating the psychometrics of the FCI assumed a unidimensional (single concept) factor structure [31, 35, 49, 67, 68, 75, 76, 85, 104, 112, 113]. These analyses found, with varying levels of confidence, that the FCI is a valid and reliable assessment.

As a result of these analyses, some items (like question 16) on the FCI have been identified as likely malfunctioning [85, 112, 113]. A malfunctioning question is one that does not function in the manner originally intended when the instrument was created. Questions can malfunction for a variety of reasons, such as wording issues, mistakes in figures, or if a question unintentionally probes a concept it was

not designed to probe. For example, question 16 on the FCI was designed to probe student understanding of Newton's third law. However, it has been found in EFA and EMIRT studies that some students are answering this question properly using Newton's first law and/or Newton's third law. As a result, this item is said to be malfunctioning since it can be correctly answered using a concept it was not originally designed to probe.

These types of item malfunctions can be detected using psychometric tools (like factor analysis and item response theory) and datasets of only correct/incorrect student responses. Datasets of this nature are said to be graded in a dichotomous manner (i.e., 1 - correct and 0 - incorrect in the dataset) and is the most common representation for student response data when analyzing conceptual assessments. Factor analysis can identify malfunctioning items by investigating whether or not items factor together in the conceptual domains they were designed to probe. If an item is found to exist in a factor whose probed conceptual domain does not match the one the item was designed for, then the item can be inferred to be malfunctioning. Similarly, an item is interpreted as malfunctioning if it is found to be too easy/hard or if the item cannot accurately differentiate between high and low performing students, both of which can be inferred using classical test theory or item response theory. Whether intentional or not, grading instruments in a dichotomous manner places all of the incorrect responses within a single question on the same level (i.e., they are all given 0's). As a result, it becomes impossible to assess the performance of the incorrect response options, often called distractors, which also play a role in the functionality of an instrument [95].

Some studies that have performed distractor-level analyses include one that used polytomous item response theory and another that examined distractor transitions matrices, both of which only investigated the FMCE [57, 88]. Polytomous item

response theory extends the models of item response theory from just correct/incorrect data to include each response option of all of the items on an assessment. Characteristics of the correct and incorrect response options can be inferred from these models, which allows for a distractor level analysis [57]. Transition matrices examine how students change their response options from pre- to postinstruciton. How students change their selections between the response options within each question helps identify changes in their thinking while considering the content of the distracting options. Except for the work presented in Chapter 2 on page 17 of this dissertation, no similar distractor-level analysis exists for the FCI.

As it currently stands, the statistical understanding of the FCI is vastly underwhelming. The factor structure of the FCI has yet to be modeled outside of exploratory analysis and the functionality of the distractors has yet to be assessed. Yet, the FCI has become one of the most commonly used instruments in PER. It is at this point in the discussion of the functionality of the FCI that this dissertation, and the work within, comes into scope. The following is a brief outline for the rest of this manuscript detailing the psychometric problems each of the chapters has attempted to answer, where references indicate original work associated with this dissertation.

To assess the functionality of the FCI at the distractor-level, Chapter 2 on page 17 presents a methodology developed by Eaton, Johnson, and Willoughby [23]. The technique presented in this chapter uses polytomous item response theory to generate partial credit scores for all of the distracting options on the FCI. The partial credit scores are given based on how the most expert-like students (i.e., students with the highest overall scores) selected individual response options. Within this framework, a question can be said to be malfunctioning in the event a distractor within the question is given a larger partial credit score than the correct response. Questions are said to be *potentially* malfunctioning if their partial credit score is

almost equivalent to the correct response's. Whether or not questions like this are actually malfunctioning would need to be decided via expert consideration of the content of the question and the partial credit scores that were given. This allows for a model-driven examination of the functionality of the distractors, and thus the questions, on the FCI.

In Chapter 3 on page 47, the question of the factor dimensionality of the FCI is examined. This is done using measures of local item independence. Specifically, statistics that measure deviations away from local item independence were used. The loss of local item independence is sometimes called local dependence. For example, local dependence between two items can appear if they i) use the same figure, ii) use the same pool of response options, iii) probe the same concept, iv) etc. Since unidimensional measurement models often assume local item independence, testing for the presence of local dependence helps to infer the dimensionality of the FCI. Specifically, this chapter seeks to address if the FCI is unidimensional or if it must be modeled in a multidimensional manner.

The conclusion of Chapter 2 is that the FCI should be examined using multidimensional models. As a result, some proposed factor models for the FCI are examined using confirmatory factor analysis (CFA) in Chapter 4 on page 89. Confirmatory FA is used to test whether a factor model that agrees with a researcher's understanding of the instrument being examined is consistent with student data. This is done by fitting a proposed factor model, either data-driven or theoretically/expert proposed, to student response data and testing for goodness-of-fit. Data-driven models generally come from previous EFA results, and theoretical/expert models result from how theory suggests or an expert thinks the items should come together in a factor structure. The three models analyzed in this chapter are i) the EFA model identified by Scott, Schumayer, and Gray [85], ii) an expert proposed model which

reflects the original intent of the creators of the FCI [38], and ii) a model developed by Eaton and Willoughby which combines the two previously mentioned models [26].

Chapter 5 on page 123 takes the results of Chapter 4 and applies them to the FCI within a confirmatory multitrait item response theory (CMIRT) framework [25]. The confirmatory part of this name comes from the pre-specification of a factor structure similar to CFA. The results from this analysis suggest that the expert models examined in Chapter 3 required minor modifications to be acceptable within a CMIRT framework. Specifically, each model requires the addition of a general trait, one which contains all of the items of the assessment. As a result, it is concluded that the FCI likely measures a primary unidimensional factor with 5-6 subfactors. That is, the FCI measures a student's overall Newtonian conceptual understanding in conjunction with their understanding of Newton's three laws, Kinematics, and Force Identification. The conceptual interpretations of the factors change slightly depending on the model being used to represent the instrument.

Lastly, in the Conclusion, the results of this dissertation are discussed, and the future of studies performed in PER are put into the context of these results.

1.1 Brief overview of item response theory and factor analysis

The following is a brief discussion of item response theory and factor analysis and is meant to give readers who are not familiar with psychometrics a conceptual understanding of how these theories function. For more technical and detailed discussions of these techniques see the references cited within the following introductory sections. As for how these tools were implemented within the studies of this dissertation, please see the methodology section within each chapter.

1.1.1 Item Response Theory

Item response theory (IRT) is a statistical method that models the interaction between a student's and an item's characteristics to predict whether or not a student will get the item correct [18]. Models that assume a student's characteristics can be represented by a single latent trait (like their ability score, θ) are referred to as unidimensional models. A latent trait is a characteristic of subjects (i.e., students) that is not directly measurable and must be inferred through related observable measures. For example, intelligence is a latent trait of humans that cannot be directly measured but can be inferred via intelligence quotient instruments. In PER, it is often said that a latent trait is a representation of a student's conceptual understanding of a specific physics subject or concept.

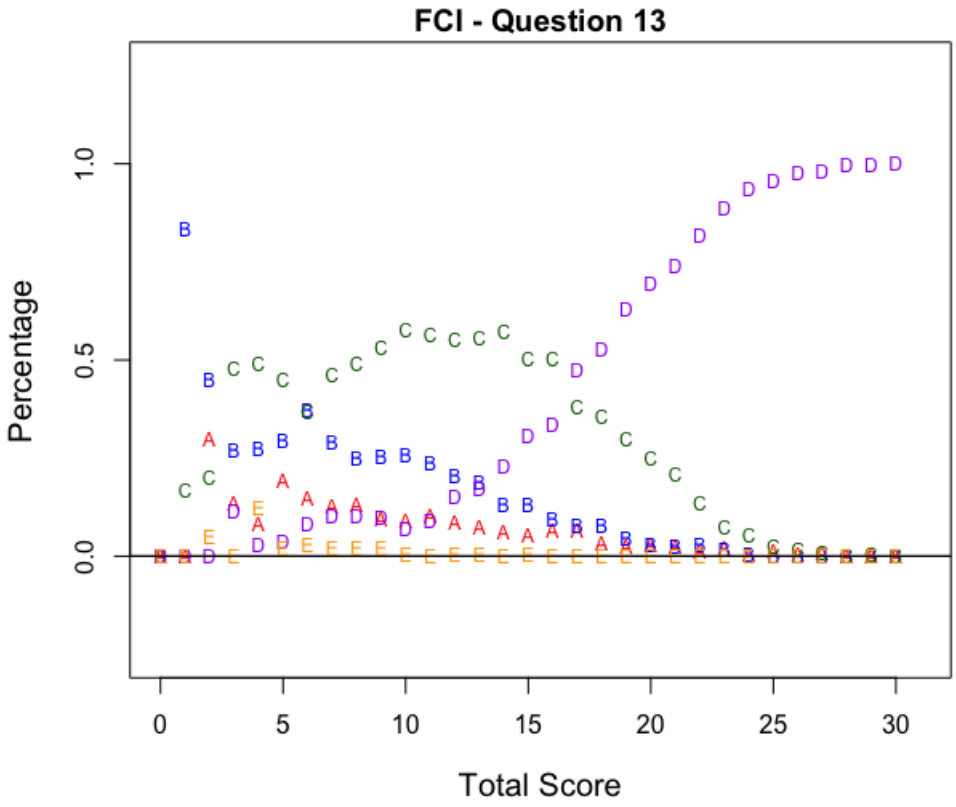
Item response theory models that assume an assessment is testing an individual student across multiple latent traits are referred to as multidimensional models. To differentiate between unidimensional and multidimensional item response theory, "IRT" is typically used to represent methods that use unidimensional models. Similarly, the name "multitrait item response theory (MIRT)" is used to represent methods that use multidimensional models.

Like most statistical methods, IRT requires two assumptions to be met before it can be used to model an assessment [56]. The first assumption asserts that the mathematical models used must adequately describe the data, and the second is that items on an assessment must be locally independent of one another. The following is a discussion of the features that an adequate IRT model must have to be viable, and a discussion of local independence is left for Chapter 3 on page 47.

First consider the item response curve (IRC) for question 13 on the FCI, Figure 1.1. An IRC is constructed by binning student responses by total score and calculating the selection percentage for each response option within each bin [67]. From Figure

1.1 it can be seen that response option D is the correct response, making the other options incorrect options (often called distractors). Since IRT only predicts whether or not a student will get an item correct, only the correct response option needs to be given a mathematical model. Modeling all of the response options for an item is referred to as Polytomous item response theory and is detailed in Chapter 2.

Figure 1.1: Item response curve for question 13 on the FCI. The horizontal axis is students' postinstruction total score and the vertical axis is the percentage of the time a specific response option was selected for each total score bin.



The IRC for question 13 suggests that the proper form of the mathematical model for the correct response option should follow a logistic curve. As a result, most of the models used in IRT are called N-parameter logistic models (NPL models). The

"N" in the previous name indicates the number of parameters being used to model an item on an assessment. So, a NPL model actually contains $N + 1$ parameters, N parameter for an item and 1 parameter for a student.

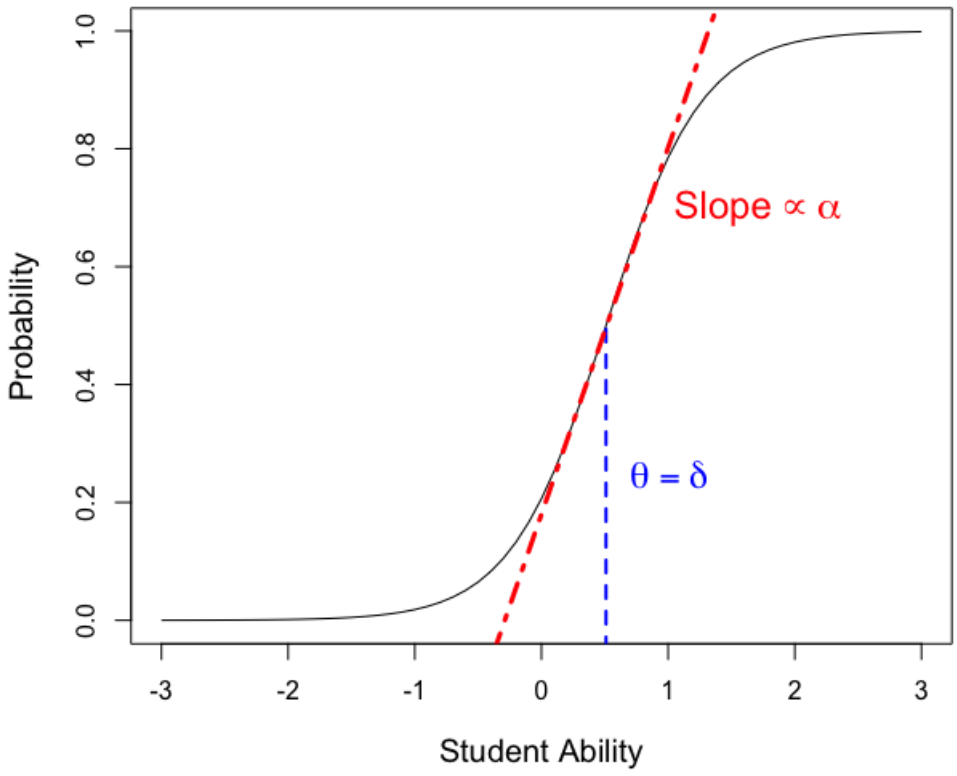
For example, the most common model used in IRT is the 2-parameter logistic model (2PL model). The typical form for this model is given as

$$P(X_i = 1|\theta, \alpha_i, \delta_i) = \frac{1}{1 + e^{-D\alpha_i(\theta - \delta_i)}}$$

where P represents a probability, $X_i = 1$ indicates this model is for answering the question correctly, θ is the student ability, α_i is the item discrimination, and δ_i is the item difficulty. In this model, the constant D is taken to be 1.702 to make the metric of the ability scale more closely relate to the traditional normal ogive metric. This implies that $\theta = 0$ is approximately the mean student ability score, and $\theta = \pm 1$ is *approximately* ± 1 standard deviation in student ability [14].

The plot of a typical result for the 2PL model can be found in Figure 1.2. Notice this model contains 3 parameters, two item parameters α and δ and one student parameter θ. Figure 1.2 shows when $\theta = \delta$ the probability of answering the question correctly is 50%. So, the larger δ is the larger a student's ability score will need to be to have a 50:50 chance of answering the question correctly. This is why δ is called the item difficulty. At $\theta = \delta$ it can be shown that the slope of the curve is proportional to α. So, larger values of α result in a steeper curve at that point, which implies that the item is better at differentiating between students who have ability scores above or below $\theta = \delta$. As a result, α is called the item discrimination to reflect its control over an item's ability to differentiate between students with ability scores above or below the item difficulty. These parameters are estimated from student response data using a maximum likelihood procedure. For more technical details on IRT and for

Figure 1.2: A typical 2PL model for an item.



more information on how the parameter estimation is performed see Refs. [4, 14].

1.1.2 Factor Analysis

Factor analysis can be broken up into two different categories, exploratory factor analysis and confirmatory factor analysis.

1.1.2.1 Exploratory factor analysis

Exploratory factor analysis (EFA) is a data-driven statistical method whose goal is to identify the underlying relationships between measured/observed variables, like scores to items on an assessment [11, 50]. Data for items being examined are taken from student responses, and as a result,

the underlying relationships discovered are potentially unique to the sample being used. Generally, this technique is used by researchers when developing scales for an instrument and/or identifying the underlying latent trait structure of the items on an assessment. A scale is a collection of items used to probe a particular latent trait domain. For example, a group of items that probe a conceptual understanding of Newton's third law would constitute a scale. This technique is often used when researchers have no a priori hypothesis about the latent trait structure for the assessment being analyzed. Additionally, EFA is generally implemented prior to applying CFA to verify that i) an underlying structure exists for the assessment, and ii) the proposed model is not radically different from the data-driven model [11, 50].

EFA is similar to the common factor model where observable variables (i.e., scores for items on an assessment) are expressed as a function of common factors, unique factors, and errors of measurement. The common factors influence multiple observed variables at a time, and the observed variables that exist within the same common factor are assumed to represent a single latent trait. The unique factors and errors of measurement account for all the information missed by the common factors. EFA is interested in uncovering the form of the common factors and how they relate to the observed variables. This structure can be found using multiple methods such as maximum likelihood or principal axis factoring. Details on how these methods are implemented can be found in Refs. [11, 50].

1.1.2.2 Confirmatory factor analysis Confirmatory factor analysis (CFA) is commonly used to test if the factor construct of an assessment is consistent with a researcher's understanding of, or intention for, the assessment. That is, CFA's goal is to fit a proposed factor model to student data [11]. The proposed models are often based on either a theoretical understanding of the content of the assessment

and/or from previous analytic research (such as a previous EFA generated model). For example, on the FCI many questions can be readily categorized by experts as probing one of Newton's three laws, Kinematics, or Force Identification. This means an expert could propose groups of questions that they, supported by their expertise in the field, think are probing the same concept. These groups of questions would form factors in the proposed factor model. A model developed like this would be said to have been generated from a theoretical perspective and not from previously performed analysis. Many methods exist for fitting a proposed model to data, all of which aim to reconstruct the item-correlation matrix. The technical specifics of how this is done is left to Ref. [11].

CHAPTER TWO

GENERATING A GROWTH-ORIENTED PARTIAL CREDIT GRADING MODEL FOR THE FORCE CONCEPT INVENTORY

Contribution of Authors and Co-Authors

Manuscript in Chapter 1

Author: Philip Eaton

Contributions: Helped develop and implement the design of this study. Developed code necessary for the project and gave the initial interpretations of the results. Wrote the first draft of the article.

Co-Author: Dr. Keith Johnson

Contributions: Helped develop and design this study and oversaw the completion of the project. Provided feedback on analysis and comments on draft of the manuscript.

Co-Author: Dr. Shannon Willoughby

Contributions: Helped develop and design this study and oversaw the completion of the project. Provided feedback on analysis and comments on draft of the manuscript.

Manuscript Information

Philip Eaton, Keith Johnson, Shannon Willoughby

Physical Review Physics Education Research

Status of Manuscript:

_____ Prepared for submission to a peer-reviewed journal
_____ Officially submitted to a peer-reviewed journal
_____ Accepted by a peer-reviewed journal
__x__ Published in a peer-reviewed journal

Published by the American Physical Society

Published on 23 December 2019

DOI: 10.1103/PhysRevPhysEducRes.15.020151

Abstract

Traditionally, multiple choice assessments are graded in a dichotomous manner, where selecting the correct option for a question awards 1 point and the selection of an incorrect option awards 0 points. As a result of this grading scheme, all incorrect response options are treated as being equally incorrect regardless of potential differences in their relative correctness, intentional or otherwise. We propose a partial credit grading model for the Force Concept Inventory (FCI) that allots points to the incorrect responses. This was done using slope parameters from the two-parameter logistics nominal response Model (2PLNRM), a model from polytomous item response theory (PIRT). The resulting scores from the partial credit model represent student growth towards a proper Newtonian mindset, as measured by the FCI. Observations indicate that this model accounts for student progression through prominent misconceptions (i.e., impetus) as their worldviews become more Newtonian. As expected, we find that student total scores increase as a result of the model, but the average overall gains on the assessment are essentially unchanged. The data used in this analysis were maintained and organized by PhysPort and included about 20 000 responses from first semester introductory physics courses at multiple universities. Ultimately, this partial credit model allows instructors to more accurately gauge the growth of their students over the course of instruction. Additionally, as a result of these partial credit scores, we are able to identify potentially malfunctioning questions on the FCI that may be sources of error in measuring student abilities.

2.1 Introduction

Traditionally, grading a multiple-choice assessment is done in a dichotomous manner, i.e. the student responses are either correct or incorrect. This method of grading places all of the incorrect options (otherwise called distractors) within an assessment on the same level of "incorrectness" by assigning them all scores of zero. However, dichotomous scoring neglects the possibility that there could exist varying degrees of incorrectness within the distractors themselves. In this paper, we propose a partial credit model that will assign nonzero values to the distractors of any given question for the well-vetted 1995, 30-question version of the Force Concept Inventory (FCI) [38], which can be found at PhysPort [64] or in Ref. [63].

Some questions on the FCI that are said to probe force identification contain

incorrect options that can readily be identified as having "levels of incorrectness" [26, 38, 82, 85]. Consider question 11 where a hockey puck slides with a constant velocity on a frictionless floor. One option states that there is only a force of gravity acting on the puck and another suggests that the force of gravity and a force in the direction of motion both act on the puck. However, the velocity of the puck is constant, so there is no force in the direction of motion. This implies that the second option could be interpreted as being less correct compared to the first. As a result, a partial credit grading model for the FCI may be appropriate to fairly assess students' understanding.

Within physics education research (PER), there have been many investigations into statistical properties of the FCI [10,24,26,31,34–36,42,49,57,68,69,75,82–86,91–93,99,104,108,112,113]. From these studies a wealth of information has been gathered about how the FCI functions and/or about how students are thinking about the FCI. Some of these studies have called into question the validity of the FCI, while others offer support to the statistical rigor of the assessment. It is appropriate to state that the FCI is not perfect. However, with more studies analyzing the statistical nature of the FCI, comes more information on how to properly create a conceptual instrument. This information is vital to the PER community when new assessments are generated in the future, which use the results of these analyses.

Partial credit scoring methods for multiple choice assessments are yet another means by which one may further understanding of student response behavior. Polytomous item response theory (PIRT) is a framework which may allow for partial credit modeling. Specifically, PIRT is a model-driven analysis which generates parameters for correct and incorrect response options, and then assigns students a latent trait ability score based on their selection of correct *and* incorrect response options [71]. Using the process proposed by Louis, Ricci and Smith [57], this study

generates a partial credit grading model for the FCI using a two-parameter logistic nominal response model (2PLNRM) from PIRT [94]. The partial credit scoring model present in this study is developed from the 2PLNRM, employed for use on the FCI with a dataset containing 19 745 student responses.

Louis, Ricci and Smith, in Ref. [57], used the 2PLNRM to propose an ordering for the relative "correctness" of the distractors within individual questions on the Force and Motion Conceptual Assessment [78]. Indicated in their study is an assumption that must be made when using the 2PLNRM in order to generate an ordering for incorrect response options:

> Students who choose correct responses on most questions are more likely to choose better (or more sophisticated) incorrect answers (for questions they got incorrect) than students who choose few correct responses.

This assumption makes conceptual sense; however, it may not give a scale of correctness in the strictest sense of the word. As will be seen in the discussion section, some orderings of distractors given by the 2PLNRM do not appear to be ordered in terms of "correctness" from an expert perspective. Instead, for some questions the ordering of the distractors appears to use alternative worldviews (i.e., misconceptions) in conjunction with a Newtonian worldview. That is, for particular questions the model assumes that the average student will go through a prominent alternative worldview, such as impetus [38, 84], before they can arrive at the desired Newtonian worldview. Thus, the resulting partial credit scores generated using the 2PLNRM inherently carry the interpretation of "progression towards proper Newtonian thinking."

A partial credit model that gives scores based on how close a student is to the proper worldview can be thought of as a growth-oriented grading model. Compared

to the stricter dichotomous scoring, partial credit scoring gives students the chance to see how much they have moved towards a proper conceptual framework. For example, a student who receives the same dichotomous total score pretest and post-test on the FCI may not actually be responding to each assessment in an identical manner. This student could have shifted from one incorrect worldview into another one while in the process of advancing towards a proper Newtonian worldview. This kind of motion between distractors is not captured in dichotomous grading, but can be reflected in a partial credit model. Thus, partial credit scores should not be thought of as an amount of correctness contained in a distractor. Instead, they should be interpreted as how close a student is to proper thinking while potentially still struggling with more complicated and harder-to-break misconceptions.

To generate a partial credit model of this nature, this study seeks to answer the following research questions:

RQ 1: What ordering for the response options on the FCI can be generated when using a teo-parameter logistic nominal response model, and as a result what partial credit grading criteria can be proposed and what information might a partial credit model reveal about the questions on the assessment?

RQ 2: How do student gains on the FCI compare between using dichotomous scoring versus the proposed partial credit scoring?

2.2 Data

The data used in this analysis were maintained and organized by PhysPort [64]. PhysPort is an online resource for physics instructors, a part of which is collecting physics education assessments and maintaining a database of student responses for many of the assessments. For this study the data originally consisted of 22 028 algebra- and calculus-based first semester introductory physics student responses,

which were matched pre-post. Students with any blank responses pre- or post-test were removed, which left a total of 19 745 student responses.

For this study, it was assumed that as students go through instruction their thinking about Newtonian mechanics becomes more coherent compared to before instruction. Thus, postinstruction student responses were used to create the partial credit grading model, since that data should be the most likely to generate a rigid and coherent model.

2.3 Methodology

First a description of the 2PLNRM will be given, followed by a discussion of how the model parameters can be interpreted. Then, the process of how to construct the partial credit model scores is detailed.

2.3.1 Two-parameter logistics nominal response model

The two-parameter logistic nominal response model (2PLNRM) was first proposed in 2010 by Suh and Bolt [94]. Their model combines the ability of a traditional two-parameter logistic model (2PLM) to model correct responses [18] and Bock's nominal response model (NRM) to model individual response options [7]. The 2PLNRM uses two different equations to model the response options on an assessment. The correct response to a question is modeled by a traditional 2PLM:

$$P(\theta) = P(x = 1|\theta, a, d) = \frac{1}{1 + e^{-(a\theta+d)}}$$

where θ is the latent ability, a is the item discrimination, and d is defined through the item difficulty, which is defined as $\delta = -d/a$. The latent ability, or student ability, is similar to the z scores of the total scores received by the students on the assessment [18]; technically, they are also parameters of the model that need to be

estimated through fitting the model to the data. Item discrimination is related to the slope of the probability curve when $\theta = \delta$. The greater the item discrimination (i.e., slope) for a given item, the more likely that a small increase in student ability could lead to a notably larger chance in getting the correct answer, and vice versa. Item difficulty is the ability associated with a 50% chance in getting the problem correct [18, 104]. For an example of the probability line generated from this model see option D (purple, dot-dashed line) in Fig. 2.1 on page 26; the item difficulty for option D is -0.194 and the item discrimination is 2.56. The "0" along the ability axis is set to be the mean of the students' ability scores (i.e., the scale is student centered).

The sum of the probabilities of a student at any ability selecting any response should be equal to 1. As such, the incorrect responses take up the remaining probability left over from the correct response's model. For m incorrect options, a single incorrect response k is modeled as:

$$\begin{aligned} P_k(\theta) &= P_k(x = 0|\theta, a, d, \boldsymbol{ak}, \boldsymbol{b}) \\ &= \left(1 - \frac{1}{1 + e^{-(a\theta+d)}}\right) \frac{e^{(ak_k\theta+b_k)}}{\sum_{i=1}^{m} e^{(ak_i\theta+b_i)}} \end{aligned}$$

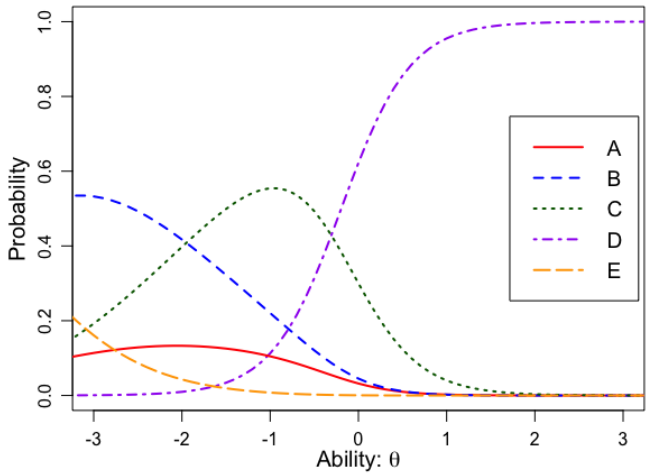
where the exponential fraction multiplying $[1 - P(\theta)]$ is the NRM for the distractors. The b_k and ak_k parameters are called the intercept and slope parameters, respectively. An example of the probability lines (also known as trace lines) can be found in Fig. 2.1 on page 26 for options A, B, C, and E. All of the item parameters and student ability estimations for this study were performed using likelihood maximization methods in the R software MIRT [14, 77]. The details of these parameter estimation techniques are outside of the scope of this study and are left to the numerous articles dedicated

to the subject, see Ref. [14] and the references cited within for more details. Since the parameters estimated in this model are generated from student responses, the resulting partial credit model should be thought of as a student-driven model. This implies that it may contain a combination of expert- and novicelike thinking, which is embodied in the assumption presented in Sec. 2.1 on page 19.

Within the NRM, slope parameters indicate how related a response option is to the latent trait being measured by the assessment. For example, if $ak_1 > ak_2$ then distractor 1 is said to be more related to the latent trait of the assessment compared to distractor 2 [8, 95]. However, slope parameters cannot be compared between different questions since the slope parameter scales are unique to each question. Thus, the ordering of the slope parameters implies an ordering of the response options within individual questions, but does not allow for direct comparisons between questions.

This interpretation of the slope parameters assumes that students who select correct options will also tend to select distractors that are more related to the latent trait than students who select fewer correct options [57]. As a result, some partial credit within questions may be allocated based on a student's progression through alternative worldviews while on a trajectory towards a Newtonian worldview (see the discussion of question 17 in Sec. 2.5.1 on page 30). Hence, the partial credit scores built from the slope parameters should be thought of as a growth and learning-based grading model, where scores may not only reflect sophistication in *actual* Newtonian thinking but may also indicate growth *towards* Newtonian thinking via motion through alternative worldviews.

Figure 2.1: Trace lines for question 13. From this plot it can be seen that option D is the correct response. The most common incorrect response is dependent on which part of ability space is being considered. The item parameters used to make this plot were taken from Table B.1 on page 195 in the Appendix. Below the figure, please find question 13 and the response options that correspond to the trace lines.



FCI 13) A boy throws a steel ball straight up. Consider the motion of the ball only after it has left the boy's hand but before it touches the ground, and assume that forces exerted by the air are negligible. For these conditions, the force(s) acting on the ball is (are):

A) a downward force of gravity along with a steadily decreasing upward force.

B) a steadily decreasing upward force from the moment it leaves the boy's hand until it reaches its highest point; on the way down there is a steadily increasing downward force of gravity as the object gets closer to earth.

C) an almost constant downward force of gravity along with an upwards force that steadily decreases until the ball reaches its highest point; on the way down there is only a constant downward force of gravity.

D) an almost constant downward force of gravity only.

E) none of the above. The ball falls back to ground because of its natural tendency to rest on the surface of the earth.

2.3.2 Generation of the partial credit scoring model

Within the 2PLNRM only the incorrect options yield NRM slope parameters because the correct answers are used to represent the underlying latent trait being measured. In the case of the FCI, the overall latent trait can be said to be "Newtonian-ness," which is simply the ability for a student to correctly answer questions on the FCI. As a result, only the incorrect responses to a question are numerically comparable in their sophistication toward measuring the latent trait. This lack of slope parameters for numerical comparisons of any correct response to its associated incorrect responses is a shortcoming of the 2PLNRM. However, a slope parameter for all correct responses may be inferred by using an inverted key.

Recall that when using the ordering proposed by the 2PLNRM under the correct key for the assessment, the response options which are the least related to the latent trait (the incorrect responses) can be quantified. An inverted key can instead be constructed and the analysis rerun using the least Newtonian options, the options least related to the original Newtonian latent trait, as given by the slope parameters, to represent the underlying trait being measured.

The latent trait examined through the 2PLNRM using the inverted key can be thought of as representing "non-Newtonian-ness" of students. Ideally, the inverted slope parameter for the correct response option, henceforth expressed as ak_{Inv}, will be identified as the least non-Newtonian option for an individual question (i.e. the most Newtonian option). This should be the case since the correct response was assumed to be the most Newtonian option under the correct key, which suggests that it should be identified as the least non-Newtonian option using the inverted key. In summary, it should be expected that the ordering of the response options under the inverted key should simply be the reverse of the ordering when using the correct key. If this is the case, the subtractive difference between the inverted key slope parameters of the

correct option and the most Newtonian distractor can be calculated. This difference can be added to the most Newtonian distractor's slope parameter under the correct key to give a measure for how much larger the correct response's slope parameter is compared to the most Newtonian distractor. This process can be summarized by using the following:

$$ak_{\text{Cor}} = (ak_{\text{Inv. 1st Dist.}} - ak_{\text{Inv. Cor.}}) + ak_{\text{1st Dist.}}$$

where $ak_{\text{1st Dist.}}$ and ak_{Cor} are the slope parameters for the most Newtonian distractor the correct response option under the correct key, respectively and $ak_{\text{Inv. 1st Dist.}}$ and $ak_{\text{Inv. Cor.}}$ are the slope parameters for the most Newtonian distractor and the correct response option under the inverted key.

Now, slope parameters can be assigned to all of the response options when using a 2PLNRM. From this ordering the assumption that the correct response option is the most Newtonian can be verified. In the case of this study all instances identified the correct option as the most Newtonian, and all of the response option orderings found using the inverted key were simply mirrored versions of the correct key's proposed ordering, as anticipated. It should be noted that some questions' correct response options were only barely identified as the most Newtonian response option. These questions will be discussed in detail in Sec. 2.5.2 on page 39.

Since the slope parameters can be thought of as being related to the correctness of response options, they can in turn be used to generate a partial credit grading model. Recall, the slope parameters for each question are on their own scales, which are unique to each question. Because of the uniqueness of these scales, the slope parameters between questions cannot be directly compared to one another. Assuming all of the correct responses are all equally Newtonian, the ak scale for each question

can be rescaled to range from 0 to 1. A value of 1 would be interpreted as being perfectly Newtonian. Partial credit scores less than 1 suggest how close a student is to a proper Newtonian worldview as measured by the question. These new scores may allow for the direct comparison of options between questions, however, this is not explored in this study and will be investigated in the future.

2.4 Results

Once the slope parameters were estimated for the entire sample, instances where options had similar or identical slope parameters were observed. To enable the statistical comparison of potentially tied options 1000 uniformly drawn classes of 10 000 students were generated and used to estimate the FCI's questions' slope parameters. The means and standard deviations of these estimations can be found in Table B.1 on page 195 in the Appendix. The root mean squared of the error of approximation (RMSEA) was calculated for each of the 1000 uniformly draw classes. The RMSEA's upper 95% confidence interval had a range of 0.0205 to 0.034 75, a mean of 0.038 58, and a standard deviation of 0.001 26. Acceptable RMSEA values are said to be below 0.06 [11, 41]. From this it can be seen that all 1000 classes had acceptable fit with the 2PLNRM.

From these random classes a distribution of ak values were found. Following Louis, Ricci, and Smith's suggestion [57], any mean ak values which were closer than 0.1 to each other were treated as being the same in the ordering of the options, see Table 2.1 on page 31. This criterion is purely *ad hoc*, functioning only to generate a visual guide for the ordering of the response options in the partial credit model and has no impact on any estimations performed. Louis, Ricci, and Smith in Ref. [57] present another method for identifying an ordering of the incorrect options using 2PLNRM slope parameters and contingency tables to establish a robust estimation

of a "hierarchy of correctness" between incorrect response options. Since the analysis method presented in this study used only the slope parameters from a 2PLNRM, and not contingency tables, a discussion of contingency tables and their uses is left to Ref. [57].

The results of the 1000 class item parameter estimations can be found in Table B.1 on page 195 in the Appendix. The proposed ordering from the ordering of question responses as given by their ak values can be found in Table 2.1 on the following page. The partial credit scores were found using the mean slope parameter values from the 1000 classes and the described methodology, see Table 2.2 on page 32.

2.5 Discussion

The following is a discussion of three questions which had the greatest subtractive difference between their average partial credits and dichotomous scores, and questions on the FCI which the partial credit model suggests may be malfunctioning. The criterion for identifying the individual questions for these discussions are elaborated upon within their respective subsections. Lastly, a discussion of the affects the partial credit model has on the total scores of students and classes are outlined.

2.5.1 3 Most changed questions

The proposed partial credit ordering of the questions, see Table 2.1 on the following page, may now be qualitatively analyzed to check for consistency. For brevity only the three questions whose partial credit and dichotomous scores have the greatest subtractive difference from one another are discussed. This difference was obtained by generating 1000 classes of 500 uniformly sampled students and comparing the mean average item scores. From this, items 14, 15, and 17 were identified as having the greatest subtractive difference between their partial credit

Table 2.1: Ordering of response options as given by the 2PLNRM. A difference of less than 0.10 in the mean slope parameter, see Table B.1 on page 195 in the Appendix, was treated as a tie in the ordering.

Questions	Ordering of response options
1:	B = E < A = D < C
2:	C < D < B = E < A
3:	E = B < D < A < C
4:	C < B < D < A < E
5:	C < E < D < A < B
6:	D < E < C < A < B
7:	D < A < C < E < B
8:	C < A < D = E < B
9:	B = D < A < C < E
10:	D < E < C < B < A
11:	B < C < E < A < D
12:	E < D < C < A < B
13:	E < B < A < C < D
14:	E < A < B < C < D
15:	E < D < B < C < A
16:	D < B < C < E < A
17:	C < E = D < A < B
18:	C < E < A = D < B
19:	B < D = A < C < E
20:	A < E < C < B < D
21:	A < B < D < C < E
22:	E < C < D < A < B
23:	E = A = D < C < B
24:	D = E < B < C < A
25:	B < A = E < D < C
26:	C < A < B < D < E
27:	E < B < A < D < C
28:	B < A < C < D < E
29:	C < A < E < D < B
30:	B < E < D < A < C

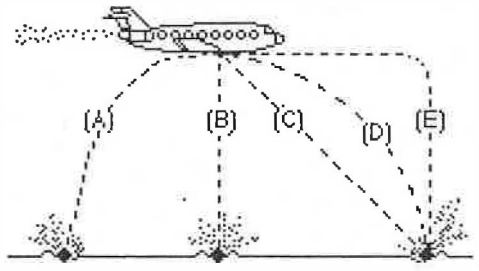
Table 2.2: The partial credit scoring model generated from the ak values given in Table B.1 on page 195 in the Appendix. Bold faced questions are ones that are potentially malfunctioning with distracting options having partial credit scores greater than 0.75, and are discussed in Sec. 2.5.2 on page 39.

Questions	A	B	C	D	E
1	0.227	0	1	0.253	0.002
2	1	0.271	0	0.164	0.308
3	0.434	0.004	1	0.084	0
4	0.662	0.320	0	0.368	1
5	0.539	1	0	0.352	0.173
6	0.433	1	0.272	0	0.155
7	0.305	1	0.469	0	0.574
8	0.565	1	0	0.613	0.616
9	0.155	0	0.428	0.05	1
10	1	0.356	0.295	0	0.098
11	0.772	0	0.261	1	0.642
12	0.548	1	0.491	0.059	0
13	0.332	0.242	0.459	1	0
14	0.575	0.686	0.776	1	0
15	1	0.529	0.859	0.116	0
16	1	0.119	0.335	0	0.947
17	0.623	1	0	0.456	0.435
18	0.258	1	0	0.276	0.119
19	0.237	0	0.330	0.199	1
20	0	0.460	0.205	1	0.089
21	0	0.208	0.579	0.394	1
22	0.562	1	0.139	0.421	0
23	0.003	1	0.268	0.008	0
24	1	0.198	0.264	0	0.013
25	0.172	0	1	0.526	0.202
26	0.293	0.371	0	0.425	1
27	0.417	0.253	1	0.546	0
28	0.159	0	0.278	0.446	1
29	0.177	1	0	0.983	0.394
30	0.496	0	1	0.339	0.165

and dichotomous scores, with Cohen's d values of 17.1, 25.2, and 21.3, respectively, on the post-test. Plots of the average items scores can be found in Figs. B.1 on page 196 and B.2 on page 197 in the Appendix.

The remainder of this section is a discussion of these three items and how their option ordering under the partial credit model aligns well with the idea of increasing Newtonian sophistication. Potential student reasoning behind the ordering of a question's options presented below is purely expert speculation. Written responses or student interviews could be used to verify that this explanation aligns with actual student thinking, which is suggested for a future study.

Figure 2.2: FCI question 14 [38]. A plane travels to the right with a velocity of $\vec{v}$ and releases a bowling ball, which it was carrying. Options A–E indicate potential paths the bowling ball could take while it fall from the plane. The proposed Newtonian ordering for this question is given as: E < A < B < C < D.



2.5.1.1 Question 14 Question 14, see Fig. 2.2, involves an airplane which is traveling to the right while carrying a bowling ball. The bowling ball is suddenly released, and the students are asked to identify the path the ball is most likely to take from five options. From the partial credit model, these options can be ordered from least to most Newtonian as E, A, B, C, and D. Option E is identified as the least Newtonian and is a "Wile E. Coyote" like path with the ball traveling to the right

without falling, then falling essentially straight down. This is the least physical option, as it is effectively impossible under any circumstance. Thus its placement as the least Newtonian option is appropriate. The remainder of the options progress as A, B, C, and then D, which from Fig. 2.2 on the preceding page appears to be an advancement of the ball's landing spot from behind (A) to underneath (B) and finally in front of the point of release (C and D). For a student to think that the bowling ball will travel backwards indicates a severe lack of Newtonian understanding; similarly, for option B. Option C is nonphysical, however, it is still more Newtonian than options A and B as the ball actually moves forwards from its point of release. Thus, the ordering of these options does agree with increasing amount of Newtonian sophistication.

2.5.1.2 Question 15 Probing Newton's third law, question 15 asks students about the relation between the force a car exerts on a truck and vice versa, see Fig. 2.3 on the next page. Since both the car and truck are accelerating to the right in this problem, students may have issues with this question by inappropriately applying Newton's second law when they should be applying Newton's third law. The ordering of the options is given as E, D, B, C, and then A, where A is the correct response. Options E and D suggest that the truck is pushed forward (to the right) simply because it is in the way of the car, and no forces are actually applied. This is clearly not the case and their positions as the least Newtonian responses is understandable. The progression from options B to C to A may be rationalized in the following manner. First, since the truck is bigger (in the figure) than the car, it will apply more force (option B). However, a discerning student may realize that the system is traveling to the right, which means the car "has" to push harder on the truck for the car and truck to accelerate to the right (option C). Finally, to get to option A the student will need to realize the forces in question are third law pairs

Figure 2.3: FCI question 15 [38]. A car is pushing on a truck and both accelerate to the right. This questions asks students about how the force from the car on the truck relates to the force from the truck on the car. The proposed Newtonian ordering for this question is given as E < D < B < C < A.

A) the amount of force with which the car pushes on the truck is equal to that with which the truck pushes back on the car.

B) the amount of force with which the car pushes on the truck is smaller than that with which the truck pushes back on the car.

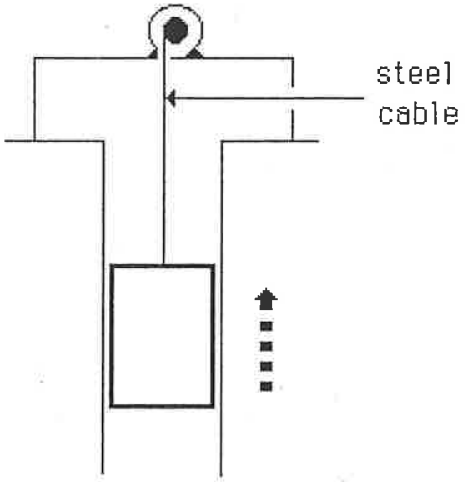
C) the amount of force with which the car pushes on the truck is greater than that with which the truck pushes back on the car.

D) the car's engine is running so the car pushes against the truck, but the truck's engine is not running so the truck cannot push back against the car. The truck is pushed forward simply because it is in the way of the car.

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

and must be equal in magnitude. Thus, this ordering of response options makes clear conceptual sense.

Figure 2.4: FCI question 17 [38]. An elevator is lifted up at a constant speed by a cable. This question asks the students about how the force exerted by the cable on the elevator is related to the force of gravity on the elevator and/or the force due to the air. The proposed Newtonian ordering for this question is given as C < E = D < A < B.



A) the upward force by the cable is greater than the downward force of gravity.

B) the upward force by the cable is equal to the downward force of gravity.

C) the upward force by the cable is smaller than the downward force of gravity.

D) the upward force by the cable is greater than the sum of the downward force of gravity and a downward force due to the air.

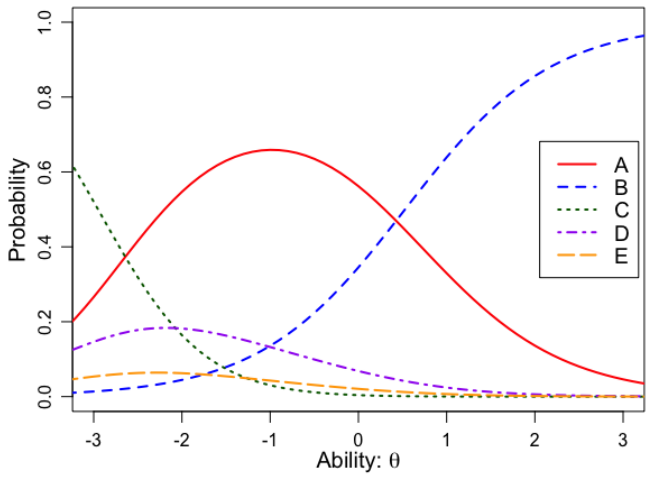
E) none of the above. (The elevator goes up because the cable is being shortened, not because an upward force is exerted on the elevator by the cable).

2.5.1.3 Question 17 Question 17 nicely exemplifies why the partial credit model is referred to as a growth-based model and not as a correctness model. This question, see Fig. 2.4 on the preceding page, shows an elevator being lifted up at a constant speed by a cable, and asks students about how the forces acting on the elevator relate to one another. The proposed ordering for this question's response options is C, E, D, A, and then B, with C being the least Newtonian. From a correctness point-of-view this ordering does not make sense, since options C and A seem equally incorrect (they both generate a net acceleration). However, when considering this model as representing growth towards a Newtonian mindset this ordering is practical.

The option ordering for question 17 suggests that when students progress from knowing little about mechanics to a Newtonian mindset they will likely move though an impetus misconception ($\vec{F} \propto \vec{v}$). For example, if option C were true then a net downward force would be generated on the elevator, which is incorrect since the elevator is moving upwards at a constant speed. However, from an impetus perspective a net force should be directed upwards as to align itself with the direction of motion regardless if the speed is constant or not, option A. Since both of these options result in a non-zero acceleration, a purely Newtonian mindset will determine these options to be equally incorrect. However, a growth-based mindset suggests that selecting option A over C indicates the student is closer to Newtonian thinking compared to selecting C over A.

This growth is also reflected in this question 17's partial credit trace lines, see Fig. 2.5 on the next page. Observing the trace lines of this question in Fig. 2.5 on the following page shows that when a student with little Newtonian ability, say $\theta = -3$, progresses towards a higher ability score, the most likely selected response option changes from option C to option A. Around $\theta = 0.5$, option B (the correct option) over takes option A and its probability of being selected grows monotonically for the

Figure 2.5: Trace lines for question 17. From this plot it can be seen that option B is the correct response, with option A being a strong distractor. The item parameters used to make this plot were taken from Table B.1 on page 195 in the Appendix.



rest of the ability space. Thus, changing from a response of C to a response of A can be interpreted as advancing toward a higher student ability score, and can thus indicate a growth towards proper Newtonian thinking.

Overall, question 17 nicely demonstrates the growth-based learning behavior of the partial credit model. Response C is being treated as the most unsophisticated as it does not represent Newtonian views nor does it align with students who still possess the impetus misconception. Response E represents students who may now begin to reflect an impetus mindset, but do not display much if any Newtonian thinking. Options D and A would both generate an upwards acceleration, which ignores the fact that the elevator is moving with a constant speed, but does make sense when viewed through an impetus lens. Option D compared to option A contains an additional incorrect force (downward force of air), and thus can be interpreted as being less sophisticated than option A based on "Newtonian-ness." Finally, the

correct response B is more likely to be chosen once impetus views are overcome and Newtonian thinking is achieved.

2.5.2 Potentially malfunctioning questions

In addition to discussing the most affected questions due to the partial credit model, it is worth discussing questions which the model suggests may be malfunctioning. In this context, malfunctioning implies that the question is not functioning as originally intended. For example, a question could be malfunctioning if students are not interpreting the written statement properly, are not selecting response options based on the expert intent reasoning, etc. Potentially malfunctioning questions can be identified by looking for large partial credit scores [95]. For instance, if all options on a question were equally spread out in terms of Newtonian-ness then the partial credit scores would come out as 0, 0.25, 0.5, 0.75, and 1. So, if a distractor has a partial credit score above 0.75, then it may be an indication that the associated question is not being interpreted properly or is malfunctioning in some other manner. Looking over the partial credit scores in Table 2.2 on page 32 shows that questions 11, 14, 15, 16, and 29 all have distractors with scores greater than 0.75. It should be noted that this method is not a "catch all and that some items which pass this investigation may still be malfunctioning. This is simply a tool that should be used in conjunction with the many other quantitative and qualitative tools used to analyze assessments.

2.5.2.1 Question 11

Question 11 asks about the forces acting on a puck which is sliding at a constant speed on frictionless ground. The response options given are lists of potential forces that may be acting on the puck. For example, the correct answer, option D, has the downward force of gravity as well as the normal force from the ground. However, option A is given a large partial credit score of 0.772. This

option offers only the downward force of gravity and does not include the normal force from the ground. Since this option excludes the normal force and does not include needless forces like options B and C, a score of 0.772 appears reasonable. As a result, question 11 is not believed to be malfunctioning, and simply contains well-separated response options in terms of their relative Newtonain-ness.

2.5.2.2 Question 14 A similar discussion can be held for option C on question 14, which received a score of 0.776. Recalling the previous discussion about this question, Fig. 2.2 on page 33 shows that option C is the only reasonable distractor since it gets the landing spot of the bowling ball correct. When compared to options A and B, the fact that path C has the ball continue in the same direction as the plane after release, and lands in a reasonable location, justifies the large partial credit score this option received. Thus, question 14 is determined to not be malfunctioning.

2.5.2.3 Question 15 Question 15 was previously discussed and has been identified as potentially malfunctioning, with option C being awarded a partial credit score of 0.859, which is generous for an incorrect response. Recall this question has a car pushing a truck to the right while both are accelerating to the right. Option C suggests that the force from the car on the truck should be greater than the force of the truck onto the car. This misconception is typically identified as either active agent (the object doing the action applied a greater force) or impetus ($\vec{F} \propto \vec{v}$) [38]. However, due to the high score of this question a more Newtonian explanation could be the issue. For example, students could be thinking in terms of Newton's second law when they should be thinking on terms of Newton's third [59]. This could lead semi-Newtonian students to select option C over the others since it appears to generate a net force in the direction of acceleration. So, option C could be being awarded points based on inappropriate utilization of Newton's second law. This suggests that

this question is not malfunctioning, and is just a difficult question for students in general. However, one could argue that it is malfunctioning, since students may not be selecting a distractor based on the intended interpretation of the distractor (i.e., using Newton's second law inappropriately versus using impetus).

2.5.2.4 Question 16 Question 16 uses the same set up and response options as question 15, however, now the car and truck are cruising at a constant speed. Because of this slight modification in the situation, option E is now awarded a partial credit score of 0.947. This question has been identified in the past as malfunctioning [85, 112, 113]. It has been observed that students are able to answer this question correctly through the use of Newton's first law, not through Newton's third [85]. If the car and truck are moving at a constant speed, and there are no resistive forces, they would not apply forces onto one another. The large score for option E corroborates these results. This question could potentially be fixed by including a statement that friction still acts on the car and truck. However, until this question is modified and reexamined it must be concluded that question 16 is malfunctioning.

2.5.2.5 Question 29 The last of the potentially malfunctioning questions to be examined is question 29, which gives option D a score of 0.983. Question 29 is a force identification question which asks about a chair sitting on the floor and offers choices for the possible forces which could be acting on the chair. These forces include: gravity, a normal force from the floor, and a downward force due to the air. Option D includes all of these forces, whereas the correct response (option C) only contains gravity and the normal force from the floor. This question was commented on by Morris *et al.* in Ref. [68] where they pointed out that question 29 "...extends beyond the Newtonian thinking dimension that the FCI is designed to measure ...". Further, in an item response theory analysis of the FCI, this question was identified as being the

least discriminating question on the FCI [104]. As a result of students' potential lack of familiarity with some of the content probed by this item, option D is awarded an unreasonable amount of partial credit. It is suggested that question 29 be examined in more depth by following the procedure used in Ref. [112], and by conducting student interviews or giving free response questionnaires for this question. Until a study of this nature is done, question 29 can only be concluded to be potentially malfunctioning for currently unknown reasons.

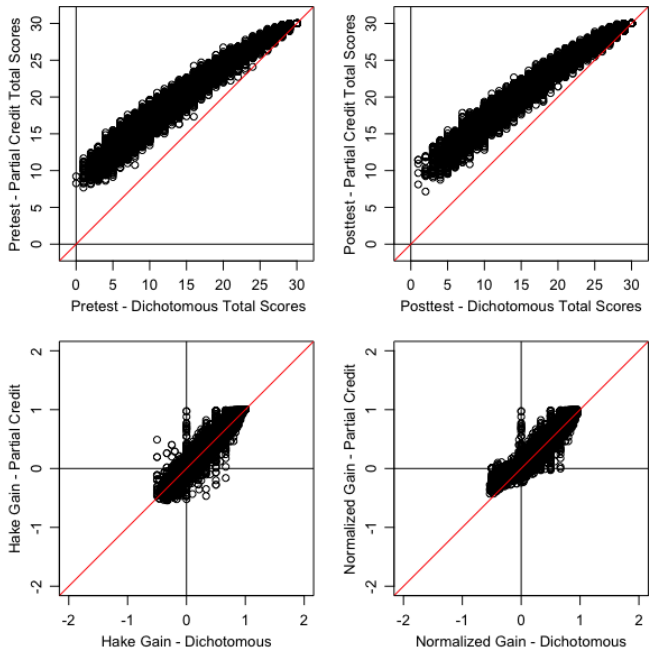
2.5.3 Comparing the dichotomous and proposed partial credit scoring models

The affect of the partial credit model on the total scores on the FCI can be found in Fig. 2.6 on the next page; the red lines are one-to-one correspondence lines. From (a) and (b) in Fig. 2.6 on the following page, it can be seen that all of the scores were improved as a result of the partial credit model. This is to be expected since the partial credit model is assigning points for responses which would have originally been given zero points. Additionally, the poorer performing students benefit more from the the partial credit compared to the top performing students. Because of the increase in students' pretest and post-test scores, concerns may be raised that an individual student's gain will change appreciably.

One gain that has been seen to be correlated to the pretest scores is the Hake gain [17, 61]. From (c) in Fig. 2.6 on the next page it can be seen that there is no drastic change to the Hake gain's values. A linear regression of the partial credit Hake gain to the dichotomous Hake gain gives a slope of 1.02 and an intercept of 0.02 with an adjusted r-squared value of 0.91. So, the average Hake gain is effectively unchanged with partial credit scoring. Similar conclusions can be made with the normalized gains [61] and can be found in (d) of Fig. 2.6 on the following page. Therefore, researchers and instructors could use the partial credit scoring presented

in place of dichotomous scoring, and should expect only small differences between the class gains found for each scoring method.

Figure 2.6: Plots comparing the results for the FCI using dichotomous scoring and the proposed partial credit scoring. (a), (b) Partial credit total scores versus the dichotomous total scores for students on the FCI pretest and post-test, respectively. (c), (d) Partial credit's versus the dichotomous Hake gain and normalized gains, respectively. The red lines in all of the plots are a slope equals 1 line to be used for reference.



However, since a unidimensional score for a multiconcept assessment is likely to be misleading for students and instructors [79], the aggregate partial credit total score should not be presented to the students. In fact, under no means should students be given their scores for an individual question in any conceptual assessment used in PER. If this kind of information became widespread these assessments would be invalidated as usable research tools. We suggest supplying the students their normalized gain using the partial credit scores, since this will be a better measure of

their Newtonian growth over the course of instruction.

2.6 Limitations and Future Works

There exist multiple methods which can be used to obtain orderings for response options, such as contingency tables [9], 2PLNRM slope parameter values, student interviews, expert orderings, etc . This study only used one of these possible methods. The other methods mentioned make different assumptions about how the ordering of response options can be obtained, which may generate slightly different orderings as a result. These other methods could be used to refine this partial credit scoring model. An example of a more rigorous treatment of finding the ordering of response options cane be found in Ref. [89] for the force and motions conceptual evaluation. Something similar to this study could be performed for the FCI, and other multiple choice conceptual assessments.

Since this study used data which came from both algebra- and calculus-based physics classes, the results may be an average of these two populations. Applying the method presented in this study to samples of just algebra- and calculus-based classes would allow for comparisons to be made between the populations which make up those courses. Any differences found in the partial credit models may illuminate differences in thought processes between the samples. Similar studies could be done separating based on gender, ethnicity, etc.

It is worth noting that this partial credit model does not fix all of the known issues the FCI has been determined to have from previous research. However, the validity of the proposed partial credit methodology is independent of the functionality of the FCI. Also, given the sensibility of the results found in this study, and the limited number of potentially malfunctioning items, we believe the presented results are reliable. If a significant number of the items had been found to be malfunctioning,

the generated model's reliability would need to be called into question.

Future work could be done by applying the methodology of this study to generate partial credit scoring models for other commonly used assessments in physics education research.

2.7 Conclusions

RQ 1: What ordering for the response options on the FCI can be generated when using a 2-parameter logistic nominal response model, and as a result what partial credit grading criteria can be proposed and what information might a partial credit model reveal about the questions on the assessment?

This study sought to create a partial credit grading model for the Force Concept Inventory. To do this a two-parameter logistic nominal response model was used to estimate the slope parameters for all of the distractors on the assessment. Using an inverted key (a key with all of the least correct responses given by the model being treated as the correct answers) gave a measure for the slope parameter for the correct responses. Under the assumption that students who get more questions correct will tend to select "more correct" distractors compared to students who got few items correct, an ordering for the response options was obtained, see Table 2.1 on page 31. The slope parameters were then adjusted to make the partial credit model by translating the slope parameter values to make the least Newtonian distractor's slope parameter zero, then rescaling so that the maximum value on the scale was 1 for all of the questions. These values can be found in Table 2.2 on page 32.

An interesting result from the presented analysis was its ability to detect potentially malfunctioning questions [95]. Questions 11, 14, 15, 16, and 29 were all found to be potentially malfunctioning, but upon investigation only questions 16 and 29 were determined as likely malfunctioning, which is in agreement with previous

research results [68, 85, 112].

RQ 2: How do student gains on the FCI compare between using dichotomous scoring versus the proposed partial credit scoring?

As one would expect, the effect of the partial credit scoring resulted in all of the students receiving higher scores on the FCI pretest and post-test compared to their dichotomous scores, with the expectation of perfect scores, see Fig. 2.6 on page 43. However, the average Hake and normalized gains were relatively unaffected. So, individual students can be given a better indication of how much they have progressed towards Newtonian thinking, while not overtly affecting the gains measured for an entire class.

As a result of this study, educators who do not know polytomous item response theory (PIRT) can now use results derived from PIRT without having to know the ins and outs of PIRT. This gives instructors the ability to assign model driven scores to distractors to give students a better sense of where they are on their path towards proper Newtonian thinking, as measured by the FCI. This would be useful in classrooms as students will be able to see their individual growth towards a more Newtonian worldview through their different selections of the distracting options from pre- to post-test on the FCI.

2.8 Acknowledgments

The authors would like to thank PhysPort for allowing us to use their data. This project was funded by the Montana State University Physics Department.

CHAPTER THREE

EXAMINING THE EFFECTS OF ITEM CHAINING IN THE FORCE CONCEPT INVENTORY AND THE FORCE AND MOTION CONCEPTUAL EVALUATION USING LOCAL ITEM DEPENDENCE

Contribution of Authors and Co-Authors

Manuscript in Chapter 2

Author: Philip Eaton

Contributions: Helped develop and implement the design of this study. Developed code necessary for the project and gave the initial interpretations of the results. Wrote the first draft of the article.

Co-Author: Barrett Frank

Contributions: Helped develop and design this study and oversaw the completion of the project. Provided feedback on analysis and comments on drafts of the manuscript.

Co-Author: Dr. Shannon Willoughby

Contributions: Helped develop and design this study and oversaw the completion of the project. Provided feedback on analysis and comments on drafts of the manuscript.

Manuscript Information

Philip Eaton, Barrett Frank, and Shannon Willoughby

Physical Review Physics Education Research

Status of Manuscript:

__x__ Prepared for submission to a peer-reviewed journal
_____ Officially submitted to a peer-reviewed journal
_____ Accepted by a peer-reviewed journal
_____ Published in a peer-reviewed journal

Abstract

Items which are chained, or blocked, together appear on many of the conceptual assessments which are utilized for physics education research (PER). However, when items are chained together there is the potential to introduce local dependence between those items. This would violate the assumption of item independence required by classical test theory (CTT), unidimensional item response theory (IRT), and other unidimensional measurement theories. Local dependence can be divided into two categories: (i) underlying local dependence (ULD), which can be adequately modeled with multidimensional measurement theories; (ii) surface local dependence (SLD), which currently cannot be modeled. The act of chaining items is thought to be one of the many potential sources of SLD between items. This study proposes two methods for detecting the local dependence and SLD which may exist between items on an assessment. These methods where then applied to the Force Concept Inventory (FCI) and the Force and Motion Conceptual Evaluation (FMCE). It was found that the assumption of item independence was violated for both assessments, implying that unidimensional measurement theories do not adequately model either the FCI or the FMCE. Further, both detection methods found a minimal amount of SLD present for FCI and a significant amount of SLD present for the FMCE. This implies that even multidimensional measurement theories are not adequate for modeling the FMCE. As a result, it is suggested that the FCI be favored until more is understood about the FMCE.

3.1 Introduction

The technique of item (question) chaining, or item blocking, is present on many of the current conceptual assessments used in physics education research (PER) [21, 37, 38, 60, 97]. Items are said to be chained together when groups of items appear in close physical proximity while probing the same concepts or when items use the same figures, response pools, reading prompts, etc. [20, 115]. This is done by test developers for numerous reasons. These reasons include, but are not limited to, less space is used to print the assessment, more items can be asked using a single figure/reading prompt/etc., and false-positive detection possibilities can be added.

The reasons for utilization of item chaining are sensible, and in some cases desirable, however they can introduce unintended statistics consequences. The most

critical issue associated with item chaining is the high potential for loss of local item independence between the chained items. Local item independence, or local independence, is the assumption that items are conditionally independent of each other, and is required for many measurement theories used in PER [18, 20]. For example, unidimensional item response theory (IRT) assumes local item independence to estimate item parameters within student response models [18]. Similarly, classical test theory (CTT) assumes the errors of items are independent from one another, which implies the items themselves are independent [20]. Thus, understanding if an assessment breaks this assumption is vital to selecting appropriate measurement theories and models.

The loss of local independence, referred to as local dependence (LD), has been investigated in the psychometric field for some time now [16, 39, 40, 55, 98, 101, 114]. However, few – if any – of these results have been implemented into assessing conceptual assessments used in PER.

In 1997, Chen and Thissen proposed separating LD into two categories: surface LD (SLD) and underlying LD (ULD) [16]. Underlying LD occurs when groups of items on an assessment share a common latent trait (e.g., a physics conception) that links items together. Effects of this nature can be modeled using multiple latent variable models, like Multi-trait IRT and factor analysis, but are not described within unidimensional models [18, 28].

Items which are chained together may possible be linked via SLD. These effects could cause students to answer items based partially, or entirely, on how they responded to the previous SLD linked items. Thus, chaining items of similar content together is a prime situation for the existence of SLD. This form of LD is problematic as most psychometric theories assume students interact with the concepts probed by each item individually. When students answer an item based on how they answered

previous items, they are not independently interacting with each concept.

An investigation into the presence of SLD between items is one way to test for the potential influences of item chaining. Since SLD is caused by a shared latent trait which links the items, resulting effects of SLD cannot be accounted for using current multidimensional models.

The first article which the authors are aware of that investigated LD between items on a conceptual assessment in PER was the the original validation of Relativity Concept Inventory's (RCI) [3]. The article discussed initial results for the RCI and identified some item pairs that were likely breaking local item independence. However, it did not go into detail about the effects that detected LD would have on parameter estimations, within IRT or CTT.

Recent efforts made towards understanding the effects of chaining items within PER conceptual assessments can be found in Refs. [93] and [106]. These studies looked into the effects of "blocking" items on the Force Concept Inventory (FCI) using multi-trait item response theory (MIRT) and modular network analysis. The methods used in these studies were then applied to the Force and Motion Conceptual Evaluation (FMCE) in Ref. [110]. It was found that the FCI's and FMCE's factor structures were likely being significantly impacted by the "blocking" of items on each assessment. However, whether these effects adversely impact the results of multiple latent variable analyses of these assessments remains unclear.

Initially, this study sought to examine the factor structure of the FMCE within an exploratory factor analysis (EFA) framework. Upon implementation of EFA, results similar to those found in Ref. [110] were obtained. It was found that the developed factor structure simply mimicked the blocked items that appear sequentially on the assessment itself. A modular network analysis of the FMCE resulted in structure which too matched the blocking structure of the assessment.

The fact that the correlational structure of items follows the chained item blocks is troubling, and lead to the investigation of the effects of chaining items.

The present study seeks to investigate the effects of item chaining on the FCI and FMCE through examination of the LD between items on each assessment. [38, 97]. The FCI and FMCE are 30-item and 47-item forced-response, multiple-choice conceptual assessments, respectively. Both of which have been used numerous times in classrooms to assess the understanding of students and/or the effectiveness of new and innovative curricula [31, 102]. Each assessment was designed to probe student understanding of Newton's Laws, with the FMCE further probing position and velocity versus time plots as well as conservation of mechanical energy. These assessments both make use of item chaining, which may affect student responses.

A clear example of this issue can be seen with items 1 and 4 on the FMCE. Both ask students to analyze a situation where a person pushes a sled across an icy surface (i.e., a frictionless surface). Item 1 prompts students to consider which force is required to push the sled to the right while speeding it up. Whereas, in item 4 the sled moves to the left while still speeding it up. Both of these items use the same response pool and figures while asking extremely similar questions; only the direction of motion changes. It is not unreasonable to infer that students could be answering item 4 based partially, or entirely, on how they responded to item 1.

As a result of item chaining of this nature, it is fair to assume that there may be a loss of local independence. If local independence is found to be broken then theories commonly used to analyze these assessment will give statistically biased results [114]. This study investigated the extent to which local independence is lost for each of these assessments, and for the affects of item chaining, by answering the following research questions:

RQ 1: By comparing simulations of local item independent assessments to the statistics of the FCI and FMCE, is the assumption of local item independence fair for each assessment?

RQ 2: By comparing simulations of ULD influenced assessments to the statistics of the FCI and FMCE, is SLD present on either of the assessments?

RQ 3: By examining the item pairs identified in research question 2, is it fair to assume that item chaining is responsible for the SLD inferred?

The rest of this work is organized as follows, first the data used in this study is specified in Sec. 3.2. Then a detailed methodology is discussed in Sec. 5.4 on page 132, followed by a presentation of the methodology's results in Sec. 3.4 on page 71. The implications of the results are considered in Sec. 3.5. Lastly the limitations of the study, a summary of the study, and suggestions for the future direction of PER can be found in Secs. 3.6 on page 84, 3.7 on page 85, and 3.8 on page 87.

3.2 Data

The data for the FCI and the FMCE came from students taking algebra- or calculus-based first year introductory mechanics (i.e., Physics I) before and after instruction had taken place. PhysPort supplied the data for both assessments [64]. The supplied data for both assessments had incomplete demographic information, so complete details of the demographic breakdown of the data is unknown.

The data for the FCI originally contained 22029 students. However, after removing any students with blank entries in their pre- or postinstruction response vectors, the sample was left with 19745 student responses. Similarly, the original FMCE data contained 19708 pre- and/or postinstruction student responses. Many of these student repressions were only for before or after instruction, not both. After

removing students that did not take both a pre- and postinstruction administration, and any student with blank responses, the FMCE sample was left with 10084 student responses. Test statistics for each of these samples can be found in Table 3.1.

The models used in this analysis required the data to be graded dichotomously. Dichotomous grading occurs when questions are sorted between two different categories for each student. In this case questions are answered either correctly or incorrectly. In the response vectors this is represented by a "1" for correct and "0" for incorrect.

Dichotomous scoring is common for the FCI, but is not recommended for the FMCE. It has been proposed that the FMCE should be graded in a blocked fashion [96, 97]. However, since this study is looking into the LD between individual items, each item will be graded separately.

3.3 Methodology

3.3.1 Item Response Theory

Item response theory (IRT) is a latent trait theory that attempts to measure students' ability scores through their interactions with items (i.e., questions) on an

Table 3.1: Test statistics for each of the samples used in this study. The data is matched pre- to postinstruction. The mean is represented by μ and the standard deviation by σ. The scores for the FMCE are calculated using the blocked grading proposed in Ref. [96].

Assessment	N	Pre-μ	Pre-σ	Post-μ	Post-σ
FCI	19745	0.437	0.213	0.608	0.221
FMCE	10084	0.317	0.244	0.537	0.296

assessment. Lord and Novick proposed two assumptions that must be met to for a mathematical IRT model to be viable [56]. The first of these assumptions pertains to the mathematical models themselves, and simply asserts that the model must describe the data well [56]. For example, a function that predicts a decreasing probability of a student answering a problem correctly as their ability increases would not match the qualitative or quantitative nature of how questions are answered. A function of this nature would break the first assumption of IRT, and would not be an adequate IRT model.

One model which satisfies this assumption is the 2-parameter logistic (2PL) model. The 2PL model returns the probability that a student with an ability of θ will respond correctly to item i given the item's discrimination index α_i and intercept index d_i. Mathematically, the 2PL model can be given as:

$$P(X_i = 1|\theta, \alpha_i, d_i) = \frac{1}{1 + e^{-D(\alpha_i\theta + d_i)}}$$

The constant D is taken to be 1.702 to make the metric of the ability scale more closely relate to the traditional normal ogive metric (i.e., $\theta = 1$ is approximately 1 standard deviation in student ability). Typically the item discrimination is factored out of the parenthesis in the exponent and then the substitution $\delta_i = -d_i/\alpha_i$ is made, where δ_i is called the item difficulty index [18]. The item difficulty index is equal to the student ability required such that the probability a student will have responded correctly is 50%. Item discrimination relates to the slope of the 2PL curve at a student ability equal to the item difficulty. Lastly, $X_i = 1$ indicates a correct response for item i, and $X_i = 0$ indicates an incorrect response.

The second assumption states that item responses are locally independent from one another, meaning students respond to an item without being influenced by the

other items on the assessment [56]. This assumption is used in parameter estimation techniques to find student ability scores and the item parameters on and for an assessment. Specifically, using the assumption of location item independence, an assessment's likelihood function can be written as the multiplicative product of all of the items' individual likelihood functions. Item and student parameters are found through maximizing the assessment's likelihood function with a given a set of student responses. This parameter estimations technique is referred to as *maximization of the likelihood*. All of the item and student parameter estimations performed in this study used the *R* package '*mirt*' [14, 77].

3.3.2 Local Dependence

The assumption of local item independence can be mathematically represented in the following manner:

$$P(X_i = 1, X_j = 1|\theta) = P(X_i = 1|\theta) \cdot P(X_j = 1|\theta)$$

This means the probability of getting items i and j correct simultaneously is equal to the probability of getting each item correct individually multiplied together. The same principle applies to getting both items incorrect, and one correct and the other incorrect. Any deviations away from this relation is an indication that students are not answering items i and j in a completely independent manner, which is to say local dependence exists between the two items.

Local dependence (LD) is separated into two categories: surface local dependence (SLD) and underlying local dependence (ULD) [16]. By definition, ULD results from unmodeled latent variables (i.e., modeling a multi-trait assessment using a single trait) which can link multiple items together [16]. This could occur on an assessment which is designed to assess multiple conceptions (e.g., Newton's Three Laws, Kinematics,

etc.). These conceptions can be thought of as being linked to a global conception (e.g., Newtonian Mechanics), but will appear in the statistics as different traits from a latent trait perspective. On the other hand, SLD occurs between pairs of items that contain highly similar content and/or are in close proximity on an assessment (e.g., chaining items of similar content together, using the same figure/response pool/reading prompt for a set of items, etc.).

Currently, it is not understood how to completely distinguish between the two types of LD when analyzing an assessment. Research into the effects of LD has been performed, and research into possibly distinguishing between the two types of LD on an assessment is ongoing [16, 40, 107, 114].

Of the two, ULD is less concerning since it can be modeled using higher dimensional models. That is, an assessment could be designed to independently measure both Kinematics and Newton's Three Laws, which would be properly described by a 4-trait multi-dimensional IRT model or a 4-trait factor analysis model [11, 18]. Alternatively, SLD can only be addressed by altering the structure of an assessment (i.e., moving items around, removing items, changing the wording of an item, etc.).

Mathematically, a simple model for SLD which links items m and n can be given

as:

$$
\begin{aligned}
&\text{With a probability of } \pi_{\text{LD}}: \\
&X_n = \begin{cases} 1, \text{ if } X_m = 1 \\ 0, \text{ if } X_m = 0 \end{cases} \\
&\text{With a probability of } 1 - \pi_{\text{LD}}: \\
&X_n = \begin{cases} 1, \text{ with } P(X_n = 1|\theta, \alpha_n, d_n) \\ 0, \text{ with } P(X_n = 0|\theta, \alpha_n, d_n) \end{cases}
\end{aligned}
\tag{3.1}
$$

where $X_m = 1/0$ and $X_n = 1/0$ are the correct/incorrect responses to item m and n respectively, and π_{LD} represents the degree, or severity, of SLD that has formed between the two items. For example, if $\pi_{\text{LD}} = 0.2$ then 20% of the time a student will answer item n based entirely on how they responded to item m. Thus, a student does not interact with item n in the manner assumed by IRT and CTT. This could occur on an assessment when multiple items that probe the same concept are asked sequentially (i.e., the items are blocked together). In this case, if the wording of the items is highly similar then students may treat all of these items as if they were a single item. This effect is apparent with items 1-7 on the FMCE, where all of these items use the same response pool and figures while probing the same conception. Situations of this nature are prime locations for students to "use the test against itself" and respond to some items while being influenced by how they answered other items in this block.

Items with close proximity and similar wording and/or content will result in a higher likelihood of developing SLD. Due to the nature in which SLD is likely to occur, this analysis can effectively identify if the act of chaining items of similar content on an assessment superficially influences how students are responding.

Another common place SLD can appear on an assessment is between the last few items. Some students will be unable to finish the assessment and will thus get these items uniformly incorrect due to not answering, and not as a result of the conceptual content of the items. So, if a group of questions at the end of an assessment is deemed to have high LD and the items do not appear to be similar in content/wording/ etc., then the LD may be a result of students simply not finishing the assessment. This particular effect will not be present within this study since only complete student response vectors were used in the analysis. Alternatively, these data could possibly be restored through the use of multiple imputations [72].

Underlying LD can be modeled via a simple bifactor model which links multiple items through a shared underlying trait. As a result of this, however, a student's apparent ability on the linked items will be a combination of their unidimensional ability and the unmodeled underlying trait ability. In the following model, θ_1^* is a student's unidimensional ability score and θ_2^* is their ability score on the underlying trait in question. The strength of the ULD between the paired items can be represented by a ULD weight, wt_{ij}. This effect can be modeled in the following manner:

$$\begin{aligned}\begin{bmatrix} \theta_{\text{item 1}} \\ \theta_{\text{item 2}} \\ \theta_{\text{item 3}} \\ \vdots \end{bmatrix} &= \begin{bmatrix} 1 & wt_{12} \\ 1 & wt_{22} \\ 1 & 0 \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} \theta_1^* \\ \theta_2^* \end{bmatrix} \\ &= \begin{bmatrix} \theta_1^* + wt_{12}\theta_2^* \\ \theta_1^* + wt_{22}\theta_2^* \\ \theta_1^* \\ \vdots \end{bmatrix}. \end{aligned} \tag{3.2}$$

The effective student ability for the item i, $\theta_{\text{item } i}$, has the potential to be significantly different from the unidimensional ability θ_1^* if the ULD weight is large. The effective student ability will be the ability with which a student will answer item i. This will artificially make the item appear more/less difficult compared to if the ULD were not present. The larger the ULD weight, the more significant the unmodeled trait is in determining how a student is responding to the items. On the other hand, if $wt_{\text{ij}} = 0$, for all i and j, then no ULD exists and the items are all locally independent, provided no SLD exists between the items. The assessment then can be assumed to be unidimensional.

From a multi-trait IRT perspective, the weights in the presented model of ULD can be interpreted as a ratio of the underlying trait item discrimination and the unidimensional trait discrimination, $wt_{\text{i2}} = \frac{\alpha_{i2}}{\alpha_{i1}}$. This can be seen in the 2PL multi-trait model,

$$\begin{aligned} P(\theta, \alpha_{i1}, \alpha_{i2}, d_i) &= \frac{1}{1 + \exp(-D(\alpha_{i1}\theta_1^* + \alpha_{i2}\theta_2^* + d_i))} \\ &= \frac{1}{1 + \exp(-D(\alpha_{i1}(\theta_1^* + \frac{\alpha_{i2}}{\alpha_{i1}}\theta_2^*) + d_i))} \\ &= \frac{1}{1 + \exp(-D(\alpha_{i1}(\theta_1^* + wt_{\text{i2}}\theta_2^*) + d_i))} \\ &= \frac{1}{1 + \exp(-D(\alpha_{i1}\theta_{\text{item i}} + d_i))}. \end{aligned} \tag{3.3}$$

where α_{i1} is the unidimensional trait's item discrimination, and α_{i2} is the ULD trait's item discrimination. Eq. 3.3 demonstrates how a multidimensional assessment could, incorrectly, be modeled unidimensionally by ignoring ULD. Consequentially, estimations of item parameters for the items influenced by the ULD will be inaccurate in a unidimensional framework.

Since student IRT ability scores are estimated using all items on an assessment, they will be relatively robust to the effects of ULD, and also SLD, provided "enough"

items on the assessment are locally independent. The more items on the assessment that are locally independent, the more robust the estimated student abilities will be as a result. Thus, ULD and SLD can be expected to significantly impact item parameter estimations, while leaving the estimated student abilities relatively unchanged [32,58].

The structure of ULD can be investigated using a multiple latent variable model like factor analysis or multi-trait IRT, which both attempt to model underlying latent trait structures. This underlying latent trait structure has been well explored for the FCI, see Refs. [26, 82, 85]. These factor models can then be assumed to represent the ULD that exists within the FCI. It should be noted that these models can be influenced by the presence of SLD, and disentangling the effects of ULD from those of SLD is not currently well understood [40].

3.3.3 Detecting Local Dependence

A common way to detect LD between a pair of items is through the utilization of contingency tables. A contingency table displays the number of occurrences for a particular combination of events. For items m and n, the contingency table records the number of times the items were answered correctly/incorrectly simultaneously or one was answered correct while the other was not. This is displayed below:

		Item n	
		0	**1**
Item m	**0**	O_{00}	O_{10}
	1	O_{01}	O_{11}

where O_{pq} is the observed number of occurrences when items m and n were answered correct/incorrect ($p = 0/1$ and $q = 0/1$).

Many useful statistics can be obtained from these tables when each element is derived from IRT-probability models. For instance the expected number of

occurrences for the contingency table above can be estimated from these probability functions. Since the parameters used in the IRT models are estimated assuming all of the items on the assessment are locally independent, any deviations between the observed and expected contingency tables can be assumed to be a result of LD between the item pairs.

Within the literature two statistics are commonly used to examine the covariation of contingency tables, Pearson's χ^2 and the logarithmic ratio G^2 statistic. Pearson's χ^2 can be calculated in the following manner:

$$\chi^2 = \sum_{p=0}^{1}\sum_{q=0}^{1}\frac{(O_{pq} - E_{pq})^2}{E_{pq}}$$

and the G^2 statistic is calculated using:

$$G^2 = -2\sum_{p=0}^{1}\sum_{q=0}^{1} O_{pq}\ln\left(\frac{E_{pq}}{O_{pq}}\right)$$

where O_{pq} and E_{pq} are the observed and expected number of occurrences from the contingency table for the pair of items being investigated [6]. Both of these statistics will depend on the size of the sample being used and compare the observed and expected number of observations from the contingency table.

Since these statistics are sample size dependent, it is often useful to employ a Cramer's V standardization to control for the sample size. This is given by:

$$V_{\chi^2} = \sqrt{\frac{\chi^2}{n(k-1)}}$$

where n is the total number of observations, and k is the number of rows in the contingency table; for this study $k = 2$. A similar expression can be written for V_{G^2}.

Each of these statistics measure variations between observed and expected values

of the contingency table. Thus, values closer to zero indicate good agreement between observations and expected results. Since the model's parameters are estimated assuming no LD, any deviations of V_{χ^2} and V_{G^2} away from 0 is an indication of potential LD linking between the items, which is unaccounted for by the model. These statistics have been found to be good indicators that LD exists [16, 39, 107, 114].

It was recently discovered that the tetrachoric correlation can be used to detect LD independent of IRT models [40]. The tetrachoric correlation is a special case of the polychoric correlation used when the sample is dichotomous. The tetrachoric correlation is calculated numerically, but can be approximated as:

$$r_{\text{tet}} \approx \cos\left(\frac{\pi}{1+\sqrt{\frac{O_{00}O_{11}}{O_{10}O_{01}}}}\right)$$

where the argument of cos is in radians. Notice, if $r_{\text{tet}} = 0$ then $O_{00}O_{11} = O_{10}O_{01}$, which implies that there was no preference to answering both items correct/incorrect simultaneously. For example, if $r_{\text{tet}} < 0$ then students tended to answer one item correct and the other incorrect, and vice versa. Other types of correlations could be used to detect LD, however the tetrachoric correlation tends to be more sensitive to correlations for dichotomous data [47].

Correlations are expected between items on assessments that probe a single concept (i.e., a unidimensional assessment). This is due to students answering items based on their latent ability, and not randomly (which would yield a correlation of zero). As a result, more difficult items will often be answered incorrectly together and visa versa for easier items. This generates nonzero correlations between the items on a single-conception assessment. When items are linked by ULD and/or SLD, the tetrachoric correlation will be artificially inflated. For this reason, item pairs with larger correlations than typical could potentially be linked via LD.

For brevity, V_{χ^2}, V_{G^2}, and the tetrachoric correlation together will be referred to as the "statistics of LD" for the remainder of the article.

3.3.4 Simulation specification

As it currently stands, no models exist that can differentiate between ULD and SLD for student responses. In order to understand the effects ULD and SLD may have on the statistics of LD, simulations using existing models were performed. This simulation methodology was proposed by Chen and Thissen in Ref. [16], and was further used by Houts and Edwards in Ref. [40] with minor variations. The main goal of these simulations was to identify whether LD is present and then how much of it can likely be modeled by ULD alone. This then allows for the testing of statistics of LD, above which ULD can no longer reasonably account for all of the LD. That is, these simulations assume that all of the LD present on a theoretical assessment is varying levels of either ULD or SLD. This allows for a comparison between the statistics of LD for these simulations and actual student responses.

In order to generate a baseline to test for LD, 200 assessments of 30-items each were generated by randomly sampling 2PL item parameters. For each generated assessment, a class of 1000 students were assigned randomly sampled latent abilities. The student and item statistics were sampled in the following manner:

- $\theta \sim$ Normal distribution(mean = 0, sd = 1)
- $\alpha \sim$ Normal distribution(mean = 1.7, sd = 0.3)
- $d \sim$ Normal distribution(mean = 0, sd = 1)

Locally independent dichotomous data was constructed for each of the randomly sampled class and assessment pairs. From these data the tetrachoric correlation matrix of the items for each simulated assessment was calculated. This tetrachoric

correlation matrix served as a baseline for comparison when testing item pairs for LD, Sec. 3.3.5 on the following page.

Item characteristics and student abilities were then estimated from the locally independent simulated data. This enabled a fair comparison of the locally independent simulations and the LD simulations (Secs. 3.3.4.1 and 3.3.4.2) while controlling for possible differences due to the numerical estimation of the item parameters. The Cramer's V standardization of Pearson's χ^2 and G^2 (V_{χ^2} and V_{G^2}) were then calculated using these estimated values.

The statistics of LD used in this study are all bivariant in nature. As such, the calculation of these statistics is independent of the number of items on an assessment [40]. This effect was tested by running the simulations for 20-item assessments and 30-item assessments, and the resulting statistics of LD were found to be independent of the number of items on an assessment. All simulation results presented in this study used 30-items for each assessment.

3.3.4.1 Surface Local Dependence Simulation To simulate SLD, items 3 and 4 were linked using Eq. 3.1 on page 58. Simulations were run for π_{LD} values that ranged from 0 to 1 in steps of 0.01. For each value of π_{LD}, 200 simulated assessments were generated using the same criteria as discussed previously, while modifying item 4's responses as per the SLD model. It is important to note that the data for item 3 remained unchanged as a result of the SLD model used for this study. Student abilities and item parameters were then estimated for each of the modified simulated data sets, and statistics of LD were calculated. This resulted in a distribution of 200 values for each of the statistics of LD for every π_{LD} value.

3.3.4.2 Underlying Local Dependence Simulation Similar to the SLD simulations, items 3 and 4 in the randomly generated assessments were treated as a pair of

items linked by ULD, as modeled by Eqs. 3.2 on page 59 and 3.3 on page 60. For simplicity, the ULD weights for each item were taken to be the same. Simulations were run for ULD weight values that ranged from 0 to 5 in steps of 0.1. For each value of ULD weight, 200 simulated assessments were generated. Student abilities and item characteristics were then estimated and statistics of LD were calculated for each assessment. This resulted in a distribution of 200 values for the statistics of LD for each ULD weight values.

3.3.5 Identifying likely LD pairs

Identification of item pairs on an assessment that are potentially linked by LD was done using two different methodologies. The first method involved the development of cutoff values for the statistics of LD found from the simulations. Stricter cutoff values were generated from the ULD simulations to test for the potential presence of SLD between a pair of items. Since the simulation results are independent of the number of items on the assessment, the cutoff values presented in Table 3.2 on page 72 can be used as they appear to test for LD and/or SLD on any assessment. This methodology requires a high level of LD between items for the pair to be flagged. The other method discussed uses one-tail t-testing to compare the simulation statistics of LD to those for an assessment. It should be noted that these methodologies yield different item pairs with the t-test method generally flagging more item pairs than the cutoff value method. This will be expanded upon in Sec. 3.4 on page 71.

3.3.5.1 Using cutoff values

A pair of items is assumed to be linked by LD if their statistics of LD are significantly larger than those of the baseline model. To test for LD, cutoff values for the statistics of LD were generated from the results of the baseline simulation. If a pair of items had statistics of LD significantly above the

generated cutoff values, then the pair was said to have LD between them. Generation of the cutoff values will be discussed below.

Items identified in this manner would violate local independence needed for IRT and CTT [18, 20]. These item pairs would thus be a source of error for any unidimensional IRT models and CTT statistics of the assessment. As a result, IRT and CTT may not accurately model the assessment and thus multidimensional models should be considered, such as MIRT and factor analysis. However, these statistical frameworks only model ULD and do not accurately model SLD [16].

Differentiating between whether the LD is caused by SLD or ULD is currently not well understood for small to moderate severity. Coincidentally, mid-range ($\pi_{\text{LD}} = 0.4 - 0.7$) and high-end ($\pi_{\text{LD}} = 0.7 - 0.8$) SLD severity result in particularly large statistics of LD. In order for ULD to result in similar statistics of LD, unusually large ULD weights must be used. These large ULD weights manifest in a 2PL model as larger slope parameters than typically found for conceptual assessments [40]. Given values of the statistics of LD, it can be inferred how likely it is for all of the LD between a pair of items to be explained with reasonable ULD weights. If it is not likely that ULD can account for all of the LD, then it can be assumed that some SLD must be present.

If it is reasonable to assume that ULD is the sole cause of any LD detected, then multiple latent trait models – like MIRT or factor analysis (FA) – can be used to properly model the assessment. Item pairs whose LD are unlikely to be a result of solely ULD implies that there is likely some SLD present for the item pair. Since SLD is not modeled in MIRT or FA, any item pairs in an assessment identified as likely possessing SLD would be a source of error. As a result, MIRT or FA would not be appropriate for the assessment.

To test for the existence of LD and to possibly distinguish between SLD and

ULD, cutoff values for the statistics of LD were proposed using the distributions from the baseline and ULD simulations. Then, these cutoff values were compared to item pair statistics of LD for both the FCI and FMCE.

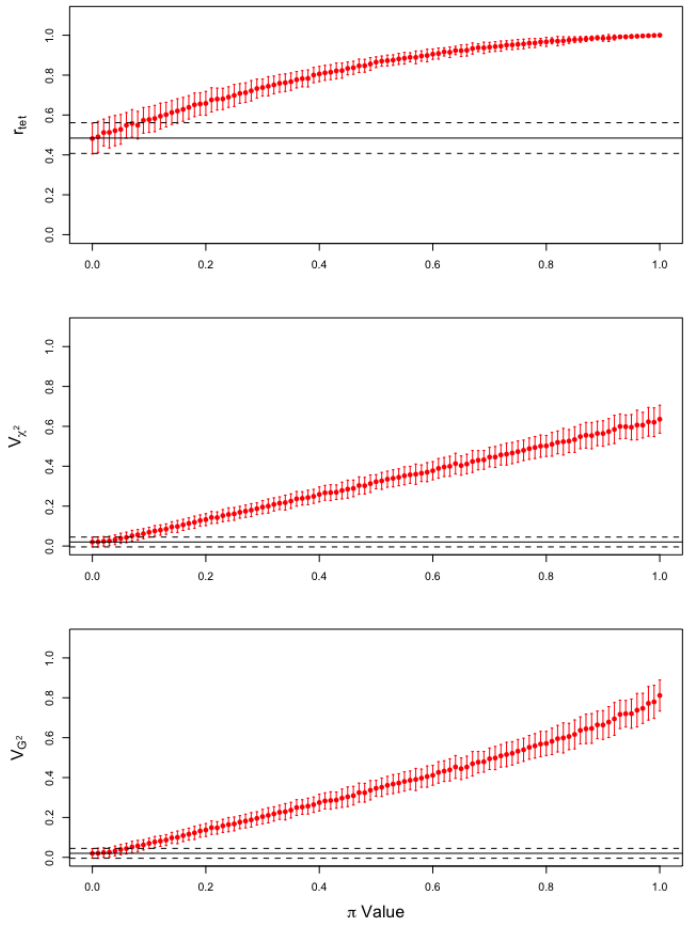
The cutoff values used in this study were taken to be the upper 95% confidence value of the statistics of LD's distributions generated by the simulations. The upper 95% confidence values are given by:

$$\text{Cutoff}(wt) = \mu_{\text{Sim}}(wt) + 1.667 * \sigma_{\text{Sim}}(wt)$$

where wt is the ULD weight value being used to generate the cutoff values, and μ_{Sim} and σ_{Sim} are the mean and standard deviation of the statistics of LD given a ULD weight. Four sets of cutoff values were generated for this study. One set was generated from the baseline simulations to test for the presence of LD, represented using $wt = 0$. Three sets of cutoff values for the statistics of LD were generated to test for SLD using different ULD weights: $wt = 1.5$, 2.0, and 2.5. These weights used for the analysis are larger than detected for typical conceptual assessments [40], but any ULD weight above 1.5 could be used to generate reasonable cutoff values to test for possible SLD. All of the cutoff values for ULD weights ranging from 0 to 5 in steps of 0.1 can be found in Table C.1 on page 199 located in the Appendix.

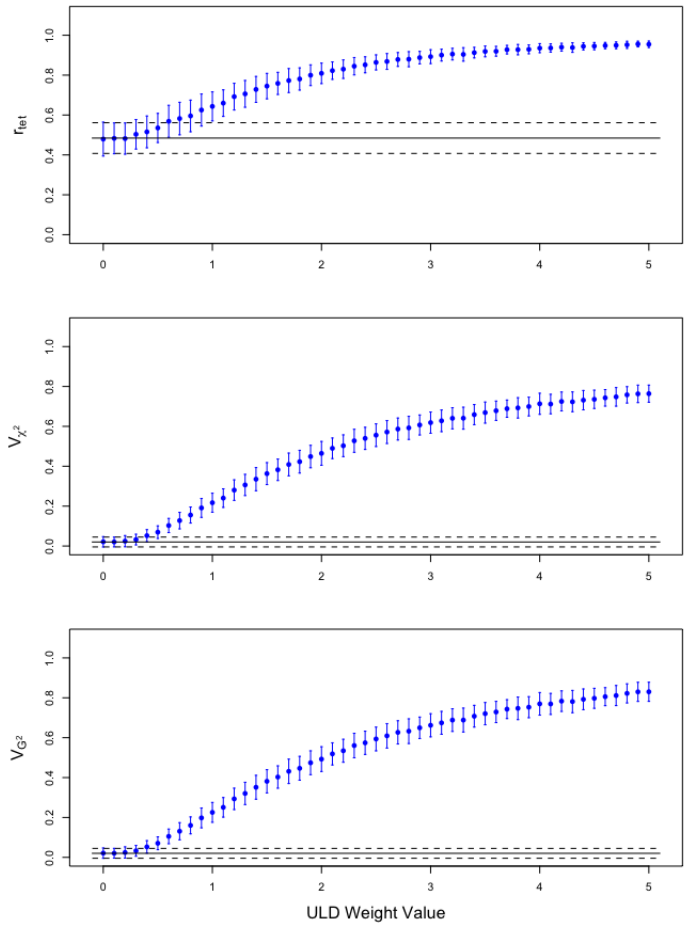
The proposed cutoff values were compared to the distributions of the statistics of LD from student data. These distributions were generated by randomly sampling 200 classes of 1000 students for both the FCI and FMCE. From here, the mean and standard deviation for each of the item pairs were calculated and representative normal distributions were used to test the significance of the proposed cutoff values. If 95% of an item pair's distribution was found to be above the corresponding cutoff value, then the item pair was concluded to likely have LD and/or SLD.

Figure 3.1: The SLD simulation's statistics of LD with versus the model's severity index π_{LD}. Mean values for the SLD simulation results are plotted as red points and error bars. The solid black lines and the dashed black lines represent the mean values and the errors, respectively. All error bars represent ± 1 standard deviation.



3.3.5.2 Using t-testing An alternative method for identifying likely pairs of items that break local independence utilized t-testing for significance. This method compares randomly sampled student responses with the simulation results via a two-sample pooled t-test [103]. To test for the presence of LD, one can use the statistics of LD that resulted from the baseline simulation. If the distributions of an item pair's statistics of LD is significantly larger than the baseline's distributions, then the item

Figure 3.2: The ULD simulation's statistics of LD with versus the model's severity index wt. Mean values for the ULD simulation results are plotted as blue points and error bars. The solid black lines and the dashed black lines represent the mean values and the errors, respectively. All error bars represent ± 1 standard deviation.



pair is likely not locally independent. The possible presence of SLD can be inferred in a similar manner by using the ULD simulations with weights larger than 1.5. All of the t-testing done in this study used $\alpha = 0.001$ for the significance level.

To reiterate, these analyses only identify pairs of items that are *likely to possess* LD and/or SLD. An item pair that is determined to likely possess SLD implies that some amount of SLD is required to explain the statistics of LD observed between

items. This does not suggest that all of the LD is accounted for by SLD alone, but that some amounts of ULD and SLD is likely.

3.4 Results

The results of the locally independent, SLD, and ULD simulations are presented below. Further, the results for testing for LD and/or SLD are presented in Sec. 3.4.2 on the next page.

3.4.1 Simulation Results

From the locally independent simulations, the Cramer's V of Pearson's χ^2 and G^2 (V_{χ^2} and V_{G^2}) were found to be identical with mean absolute value of 0.02 and standard deviation of 0.025. Further, the tetrachoric correlation was found to have a mean absolute value of 0.484 with a standard deviation of 0.077. The effect of varying SLD and ULD severity on the statistics of LD is shown in the top and bottom plots, respectively, of Figures 3.1 on page 69 and 3.2 on the previous page. These figures show that the statistics of LD deviate significantly from locally independent results at relatively small amounts of LD. For SLD this occurs at around $\pi_{\text{LD}} = 0.2$ for all three of the statistics of LD. Alternatively, for ULD this occurs at a ULD weight of about 0.75 for V_{χ^2} and V_{G^2} and 1.25 for the tetrachoric correlation.

From the plots in Figures 3.1 on page 69 and 3.2 on the previous page it can be seen that the tetrachoric correlation has large standard errors for small amounts of ULD and SLD, and get progressively smaller as the severity of ULD and SLD increase. Alternatively, the standard errors of the Cramer's V of Pearson's χ^2 and G^2 begin small and get larger with the severity of ULD and SLD increase. This implies the tetrachoric correlation will likely identify fewer linked items for low ULD and SLD, where V_{χ^2} and V_{G^2} are more precise and will be able to identify more item pairs.

Thus, the tetrachoric correlation will identify more item pairs at higher LD severity and the V_{χ^2} and V_{G^2} statistics will identify more item pairs at lower LD severity.

It should be noted that the plots of the affects of SLD and ULD can not be directly compared to one another since the horizontal axes are not on the same scale. Further, these scales cannot be mapped to the same scale in any meaningful manner.

3.4.2 Identifying likely LD pairs

Presented below are the pairs of items flagged by the cutoff and t-testing methodologies for the FCI and FMCE. Item pairs were first tested for LD then subsequently tested for SLD.

Table 3.2: The proposed cutoff values for the statistics of LD for some given ULD weight values. The cutoff values for a ULD of "0" was used to detect the presence of general LD within an item pair.

ULD weight	0	1.5	2.0	2.5
r_{tet}	0.613	0.851	0.895	0.927
V_{χ^2}	0.060	0.455	0.564	0.650
V_{G^2}	0.060	0.478	0.596	0.693

3.4.2.1 Using cutoff values

Proposed cutoff values for the statistics of LD were generated using the procedure described in Sec. 5.4 on page 132. For reference, the cutoff values from the locally independent simulation can be found in the column labeled "0" within Table 3.2. These results are independent of the total number of items, so the cutoff values in Table 3.2 can be used to test for the presence of LD in any assessment. For brevity, the results for the three individual statistics of LD will

Table 3.3: The pairs of items identified as possibly linked by SLD on the FMCE via the cutoff values method. The numbers in the table separated by ":" indicate the question pairs whose statistics of LD significantly larger than the cutoff. Items listed in regular text indicated pairs detected on both the pre- and postinstruction assessments, items in parenthesis indicate pairs flagged only on the preinstruction assessment, and bold for only the postinstruction assessment.

Cramer's V of Pearson's χ^2						
ULD weight	1.5			2.0		2.5
	32:34		36:38	**36:38**		
Cramer's V of G^2						
ULD weight	1.5			2.0		2.5
	32:34		**36:38**	**36:38**		
Tetrachoric Correlation						
ULD weight	1.5			2.0		2.5
	1:2	1:4	**3:7**	**1:2**	1:4	14:17
	8:9	**8:10**	8:11	**8:9**	(11:12)	(24:26)
	9:12	11:12	11:13	**11:13**	**14:16**	**36:38**
	14:16	14:17	14:18	14:17	16:18	
	14:19	16:17	16:18	16:19	22:23	
	16:19	17:18	(17:19)	24:26	(27:28)	
	18:19	22:23	(22:26)	**27:29**	32:34	
	24:26	27:28	**27:29**	**36:38**		
	32:34	36:38	**46:47**			

be given in a

(V_{χ^2} results, V_{G^2} results, r_{tet} results)

format.

The postinstruction FCI was found to possess (17, 18, 9) pairs of items that are not locally independent, involving a total of (17, 18, 11) individual items. This represents 2% - 4% of the total item pairs possible for the FCI and involves 30% - 60% of the individual items. The preinstruction FCI was found to contain (21, 21, 8) pairs of items as not being locally independent, including (18, 18, 10) individual items. This accounts for 1.8% - 5% of the total item pairs on the FCI and 33.33% - 60% of the total items.

Table 3.4: The pairs of items identified as possibly linked by SLD on the FMCE via the t-testing method. The numbers in the table separated by ":" indicate the question pairs whose statistics of LD significantly larger than the cutoff. Items listed in regular text indicated pairs detected on both the pre- and postinstruction assessments, items in parenthesis indicate pairs flagged only on the preinstruction assessment, and bold for only the postinstruction assessment. All of the found pairs have a significance of $p < 0.001$.

Cramer's V of Pearson's χ^2 and G^2				
ULD weight	1.5		2.0	2.5
	(1:4)	**30:32**	32:34	**36:38**
	30:34	32:34	36:38	
	36:38	46:47		

The postinstruction FMCE results found (249, 256, 222) item pairs as not being locally independent, involving a total of (45, 42, 45) individual items. These results account for 21% - 24% of the total number of item pairs on the FMCE and 90% - 96% of the individual items. The preinstruction results for the FMCE were more severe with (386, 390, 203) item pairs being identified as losing local item independence. This

Table 3.5: The pairs of items flagged on the FMCE and FCI as likely being linked with some SLD via the polychoric correlation and the t-testing method. All of the item number listed were found to have paired with the items number of the far left, for the varying ULD weight values. Items listed in regular text indicated pairs detected on both the pre- and postinstruction assessments, items in parenthesis indicate pairs flagged only on the preinstruction assessment, and bold for only the postinstruction assessment. All of the found pairs have a significance of $p < 0.001$.

FMCE - Tetrachoric Correlation			
	ULD weight values		
Item	1.5	2.0	2.5
1	2, **3**, 4, 14, 16, (17), 18, 19	2, 4	**2**, 4
2	1, **3**, 4, 14, 16, 17, 18, 19	1, **4**, **14**	**1**
3	**1**, **2**, **4**, **6**, 7	7	**7**
4	1, 2, **3**, 14, 16, 18, 19	1, **2**	1
6	**3**, **7**		
7	3, **6**	3	**3**
8	9, 10, 11, 12, 13, (14), (16), 18, 21, 27, (28), (29)	9, 10, 11, 12, **13**	9, **10**, 11
9	8, 10, 11, 12, **13**, 28	8, **10**, 11, 12	8, 12
10	8, 9, **11**, 13	8, **9**, **13**	**8**
11	8, 9, **10**, 12, 13, (18), **21**, 27, 28, **29**	8, 9, 12, 13, **27**	8, 12, 13
12	8, 9, 11, 13, **27**, 28	8, 9, 11, **13**, **28**	9, 11
13	8, **9**, 10, 11, 12, **27**, **29**	**8**, **10**, 11, **12**	11
14	1, 2, 4, (8), 16, 17, 18, 19, (20), **23**, **24**, **25**, **26**	**2**, 16, 17, 18, 19	16, 17, 18, 19
16	1, 2, 4, (8), 14, 17, 18, 19, 20, **23**, **25**	14, 17, 18, 19	14, 17, 18, 19
17	(1), 2, 14, 16, 18, 19, (20), **23**, **24**, **26**	14, 16, 18, 19	14, 16, 18, (19)
18	1, 2, 4, 8, (11), 14, 16, 17, 19, 20, 21, **22**, 23, **24**, **25**	14, 16, 17, 19, (20), **23**	14, 16, 17, 19
19	1, 2, 4, 14, 16, 17, 18, 20, **23**, **25**	14, 16, 17, 18	14, 16, (17), 18
20	(14), 16, (17), 18, 19	(18)	
21	8, **11**, 18		
22	**18**, 23, 24, 25, 26	23, 24, 25, 26	23, (26)
23	**14**, **16**, **17**, 18, **19**, 22, 24, 25, 26	**18**, 22, 24, 26	22
24	**14**, **17**, **18**, 22, 23, 25, 26	22, 23, **25**, 26	26
25	**14**, **16**, **18**, **19**, 22, 23, 24, 26	22, **24**	
26	**14**, **17**, 22, 23, 24, 25	22, 23, 24	(22), 24
27	8, 11, **12**, **13**, 28, 29	**11**, 28, 29	28, **29**
28	(8), 9, 11, 12, 27, 29	**12**, 27, 29	27, (29)
29	**11**, **13**, 27, 28	27, 28	**27**, (28)
30	**31**, 32, 34	**32**, 34	
31	**30**, **32**, **34**		
32	30, **31**, 34	**30**, 34	34
34	30, **31**, 32	30, 32	32

Table 2.5 continued...

FMCE - Tetrachoric Correlation

	ULD weight values		
Item	1.5	2.0	2.5
36	38	38	38
38	36	36	36
40	42		
42	40		
44	**45**		
45	**44**		
46	47	**47**	**47**
47	46	**46**	**46**

FCI - Tetrachoric Correlation

	ULD weight values		
Item	1.5	2.0	2.5
5	18		
18	5		
25	26		
26	25		

involved (47, 47, 40) individual items, representing 19% - 30% of the total possible item pairs on the FMCE and 85% - 100% of the items. For a summary of these results for the FCI and the FCME see Table 3.6 on the next page.

From these results it can be seen that both assessments contain LD between items. This is expected since both assessments were originally constructed to measure multiple conceptions. Provided no SLD is present on either assessment, they both can be properly modeled using a multiple latent variable theory (such as MIRT or FA).

To test for the possible presence of SLD, cutoff values were generated using ULD weights of 1.5, 2.0, and 2.5, which can be found in the last three columns of Table 3.2 on page 72. In conjunction with the distributions of the statistics of LD from the student data, these cutoff values were used to identify item pairs where the LD is

Table 3.6: The number of item pairs identified via t-testing as potentially breaking local item independence for the FCI and FMCE pre- and postinstruction. The N column represents the number of item pairs detected. The % column denotes the percentage of the total item pairs made up by the values in the N column.

	V_{χ^2}		V_{G^2}		r_{tet}	
Assessment	N	%	N	%	N	%
Pre-FCI	21	2.4	21	2.4	8	0.92
Post-FCI	17	2.0	18	2.1	11	1.3
Pre-FMCE	386	17.9	390	18.0	203	9.4
Post-FMCE	249	11.5	256	11.8	222	10.3

unlikely to be explained solely by ULD.

For the FCI, zero item pairs were flagged for potentially being linked through SLD for both pre- and postinstruction administrations. This implies that the detected LD on the FCI is likely a result of only ULD, which can be modeled using a multiple latent variable model. Thus, the act of chaining items on the FCI is not significantly effecting how students respond to items. Note that this result disagrees with previous literature, see Ref. [93] which found that item blocking was significantly effecting the results for the FCI. This suggests that more research should be performed to fully understand the effects item chaining has on this instrument.

The FMCE had many pairs of items flagged as potentially being linked through SLD. That is, the likelihood that ULD alone can account for the observed statistics of LD is very low, and thus there is likely a combination of both ULD and SLD linking the items. Table 3.3 on page 73 shows the items flagged by the cutoff value methodology for the FMCE.

For the postinstruction data, (2, 2, 25) pairs of items were identified as possibly being linked through SLD when using the smallest ULD weight ($wt = 1.5$). When considering the larger ULD weights, (1, 1, 13) and (0, 0, 2) item pairs were identified, see Table 3.3 on page 73. For the preinstruction data, (1, 1, 21) pairs of items were identified as possibly being linked through SLD when using the smallest ULD weight. When considering the larger ULD weights, (0, 0, 9) and (0, 0, 2) item pairs were identified.

3.4.2.2 Using t-testing The results presented using the cutoff values method were sufficient for revealing that both the FCI and FMCE do not posses local item independence. Consequentially, the results of the t-test method adds little extra information to what has already been revealed about the existence of LD. The results presented below for the t-testing method only investigated the possible presence of SLD.

For the FCI, two pairs of items were identified as possibly being linked by SLD, items 5:18 and items 25:26. These item pairs were flagged using a ULD weight of 1.5 and the tetrachoric correlation for pre- and postinstruction data. None of the other ULD weights or statistics of LD flagged any item pairs, pre- or postinstruction. It can be inferred that either the FCI has a small amount of SLD or these items are linked through a strong underlying trait.

The results of the t-test analysis for the FMCE can be found in Tables 3.4 on page 74 and 3.5 on page 75. The item pairs identified using the Cramer's V of Pearson's χ^2 and G^2 were identical. These statistics flagged 5 pairs, 2 pairs, and 1 pair of items when using ULD weights of 1.5, 2.0, and 2.5, respectively. The tetrachoric correlation flagged 78, 36, and 23 pairs of items for the ULD weights postinstruction. Similarly, for the preinstruction data 77, 35, and 23 item pairs were flagged. Due to

the number of flagged item pairs on the FMCE, it is unlikely that all pairs can be explained via strong underlying traits alone.

3.5 Discussion

Discussed below is the likelihood that SLD invalidates multivariate models for the FCI and FMCE. Further a discussion of unidimensional scoring and possible solutions is presented.

3.5.1 Multivariate Models of the FCI and FMCE

As was presented previously in the Sec. 3.4 on page 71, most of the LD flagged item pairs on each assessment can be explained using only ULD.

For the FCI only two item pairs were flagged as potentially being linked, in part, via SLD. These were found using only the most lenient of testing criteria presented in this article. Thus, it can be assumed that these items are linked either by small π_{LD} values, or by a very strong underlying trait. This implies that the error introduced to a multiple latent variable model of the FCI by these item pairs will likely be small. Researchers concerned about this error can simply remove either item 5 or 18 and remove either 25 or 26.

Of the two instruments, the FMCE was found to contain far more SLD. For the most lenient testing criteria, 78 pairs of items were flagged as likely containing some amount of SLD (compared to the two found for the FCI). As a result, multiple latent variable models that assume local item independence may not accurately represent the FMCE.

Recall, in some cases item chaining is the practice of using the same figures, response pools, reading prompts, etc. for groups of items [115]. From this it can be seen that of the two item pairs flagged for the FCI, only the item pair 25:26 actually

meets this criteria. Thus, only this item pair on the FCI can be assumed to be impacted by item chaining. Whereas, on the FMCE 78 item pairs were identified as likely being linked in part by SLD. Of these 78 identified items, more than half meet the criteria of being chained/blocked items. This implies that item chaining is having a significant impact on how the items on the FMCE are functioning.

In a previous study of the FMCE, Yang, Zabriskie, and Stewart obtained exploratory FA and MIRT models [110]. Since, both exploratory FA and MIRT assume only ULD exists between items, the possible presence of SLD may significantly impact the results of these methodologies. In fact, a comparison of the flagged item pairs in Table 3.5 on page 75 and the results of the factor analysis presented in Ref. [110] reveals that many of the factors identified may be linked through a combination of ULD and SLD. Similarly, many of the major links present in the partial correlation networks are flagged in this analysis as likely containing some level of SLD. This however, does not imply that the results presented in Ref. [110] are incorrect. The results of this study imply that some of the observed correlations used to generate the factor models and network structures are likely being artificially inflated due to SLD. To test the validity of the proposed models, further exploration into the affects of SLD on multivariate models is recommended.

3.5.2 Scoring the FCI and FMCE

Both assessments were found to violate the assumption of local item independence. This implies that unidimensional models, which assume local item independence, should not be used to analyze or score either of these assessments. The effects of LD on the results of unidimensional IRT models have been explored in detail, see Ref. [115]. Within, Yen details the effect LD can have on total scores, assessment validity, IRT test information, and IRT item parameter estimations. The

exact details of these effects are outside the scope of this study. From the results presented in Ref. [115] it can be inferred that unidimensional IRT will not be accurate if LD is present in an assessment.

Similarly, CTT statistics are affected by the presence of LD within an assessment. For example, the effects of SLD as a proxy of LD on classical item difficulty and discrimination are shown in Figures 3.3 on page 83 and 3.4 on page 84. Within each figure the SLD influenced values are plotted versus the original locally item independent values for varying severity of SLD. Details concerning CTT statistics can be found in Ref. [48]. From these figures it is apparent that as SLD severity increases the induced error of observed classical measures also increases. Due to the detection of potential SLD for the item pairs listed in Tables 3.3 on page 73, 3.4 on page 74, and 3.5 on page 75 it can be assumed that the CTT statistics measured for these items are likely inaccurate. It can then be inferred that aggregate total scores made up of SLD linked item will not accurately reflect student knowledge. As these preliminary results show, the presence of LD on an assessment can drastically impact the observed CTT statistics for an assessment.

To address the possible effects of LD, Yen suggests using locally independent testlets (grouped items that are graded together) in place of locally dependent item blocks [115]. Testlets can be formed by grouping items that share LD, then a grade can be assigned to each testlet individually. If each of the testlets are locally independent, then IRT can be preformed by treating each of the individual testlets as "items". This would result in a reliable measure of student ability, while controlling for the effects of LD.

Testlets for the FCI could be formed using one of the three models tested in Ref. [26] where each testlet would contain a single factor. This would result in students receiving one score for their understanding of each Newtonian concept represented

by these models. These testlets could then be used to generate an IRT model of the FCI which would supply a student with their "Newtonian" ability.

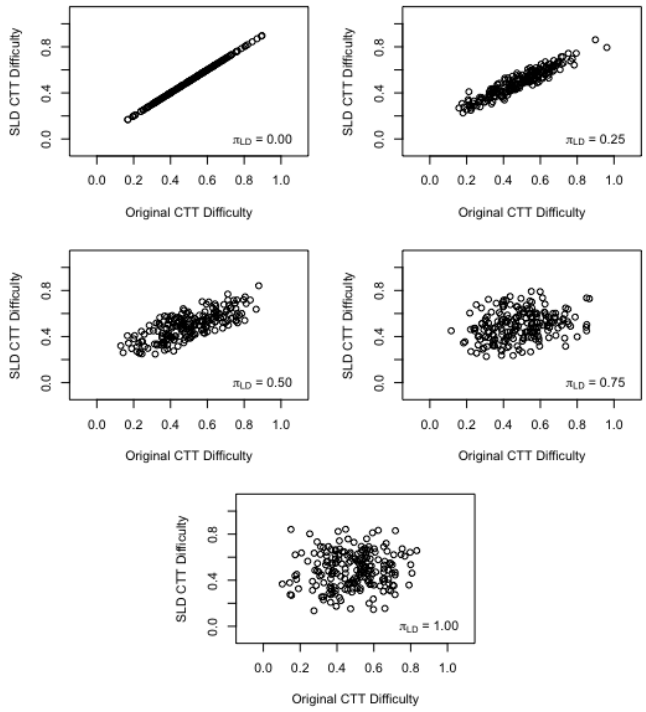
Due to the potential presence of SLD on the FMCE multiple scoring options may be considered: (1) forming testlets from LD linked items, (2) rearranging the items which appear on the FMCE, and (3) splitting the FMCE into two smaller assessments. Each of these methods contain drawbacks, and it is possible that the best option is to simply rewrite the instrument.

The testlet scheme for the FMCE proposed by Thornton *et al.* is well in line with Yen's suggestion, however some of the suggested testlets are not locally independent and thus do not meet all the specifications indicated by Yen [96, 115]. The required modifications to ensure these testlets are locally independent would result in the FMCE being made up of one dominant testlet and many smaller testlets.

Rearranging the items on the FMCE would entail generating unique physical descriptions, response options, and figures for each of the items on the assessment. The separated items could then be randomized to ensure students are being primed for concepts as minimally as possible. The resulting assessment would need to undergo extensive analysis to fully elucidate its statistical properties. However, considering the initial reasons for chaining the items, and the added false positive detection benefits, this "fix" may make the assessment something completely different than what the creators of the FMCE originally intended [90, 96, 97].

The splitting of FMCE could be done in a similar manner to how both the FCI and the Conceptual Survey of Electricity and Magnetism were separated in Ref. [33, 109]. However, due to the potential SLD found in this study, any estimated characteristics for each individual item are likely not accurate, and will need to be reexamined after any of these suggested changes are made. It should be noted that these new assessments may not contain the same false positive detection abilities as

Figure 3.3: Plots of the classical test theory item difficult indices of the SLD modified results versus the original item difficulties. Beginning from the top left, these plots are for $\pi_{\mathrm{LD}} = 0.00$, 0.25, 0.50, 0.75, and 1.00. Notice as π_{LD} gets larger, the classical difficulty becomes increasingly affected.

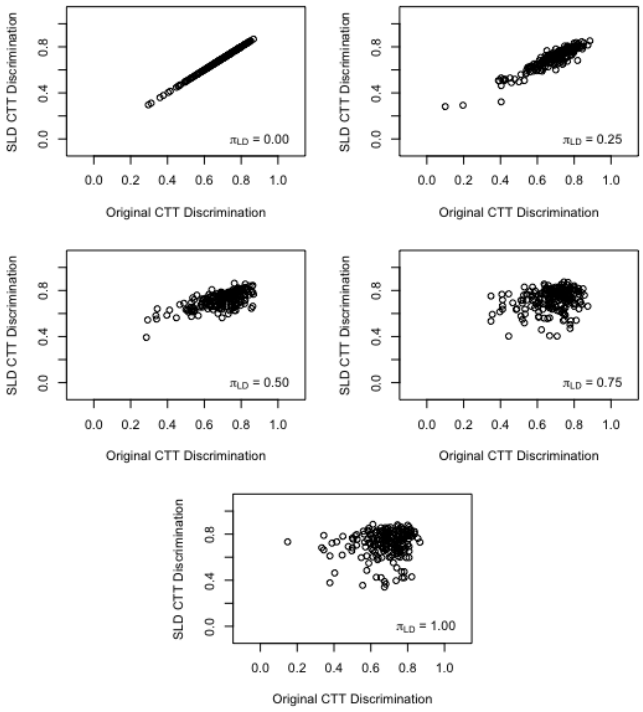


the original form of the FMCE. This would suggest that splitting the assessment in half may not be recommended, see Ref. [90].

Ultimately, a single-number grading scheme could be readily created for the FCI using current research results. However, given the extent of LD found and the large possibility of SLD being present, the FMCE should not be graded using a single number until a new grading scheme has been proposed and studied.

Considering the LD present on each of the assessments, only multivariate models should be used to assign scores for either of the assessments. Until the SLD which is likely present on the FMCE is better understood, any multi-dimensional measures

Figure 3.4: Plots of the classical test theory item discrimination indices of the SLD modified results versus the original item discrimination. Beginning from the top left, these plots are for $\pi_{\mathrm{LD}} = 0.00$, 0.25, 0.50, 0.75, and 1.00. Notice as π_{LD} gets larger, the classical discrimination becomes increasingly affected.



for the FMCE should be treated as being inaccurate. Thus, the results of this study suggest that the FCI, and not the FMCE, be used in conjunction with a multi-dimensional model to probe the Newtonian understanding of students.

3.6 Limitations

The data used in this study was a mixture of algebra- and calculus-based introductory physics classes. Some of the LD found in this study may belong more to one of these groups over the other. However, since the FCI and FMCE are intended to be used for both courses, the presence of potential LD suggests that both assessments

should be modeled multi-dimensionally. Despite this, the potential presence of SLD on the FMCE is alarming and should prevent its further use until more studies have investigated the possible SLD detected here.

The interpretations made in this study assumed that LD was only generated through ULD and SLD. There could be effects other than ULD and SLD which could have generated some of the LD present on the FCI and FMCE. As it currently stands, the literature into LD does not offer any other classifications aside from ULD and SLD. The interpretations made in this study ignored the possibility of LD sources other than ULD and SLD.

The models used for SLD and ULD are extremely simplistic in nature and may not fully capture the effects. This may be particularly true for models that contain more traits, and thus better explain the ULD possibly linking items. In a future study, simulations will be performed using FMCE specific models to better replicate the ULD present, such as the MIRT model presented in Ref. [110]. Any unexplained LD could then be interpreted as potential SLD. However, multiple models would need to be used to ensure the unexplained LD is likely SLD and not by unmodeled ULD.

3.7 Summary

The FCI and FMCE were examined to determine the extent to which local independence was broken on each assessment. Local dependence occurs when multiple items influence one another in a manner that prevents students from responding to the items as though they were independent. Chen and Thissen [16] differentiated the causes of LD by defining two categories: surface local dependence and underlying local dependence.

When multiple items on an assessment share a common conception (or trait), it is said that ULD is linking the items. Since unidimensional IRT does not model

multiple traits, its results are affected by ULD. Conveniently, the effects of ULD on an assessment can be accounted for by using multidimensional models when analyzing the assessment, such as factor analysis and multi-trait item response theory.

Items which share common wording, figures, answer banks, reading passages, etc. or which are chained (blocked) together are likely linked via SLD. These effects cause students to answer items based entirely on how they responded to the previous SLD linked items. As a result, students do not interact with items independently, which is assumed by IRT and CTT. Since SLD is not due to a shared latent trait between the items, the effects of SLD cannot be accounted for by using multidimensional models. If an assessment contains SLD then all statistics and models related to the assessment should be called into question for correctness and validity. For this reason, SLD is more concerning than ULD.

The assessments in question were found to contain item pairs which break local item independence. Of the items on the FMCE, anywhere from 85% to 100% where found to not be locally independent. On the other hand, 30% to 60% of the items on the FCI were found to lack local independence. As follows, results of unidimensional models for the FMCE will likely contain a significant amount of error. Models for the FCI, on the contrary, will likely contain some error, but not as much as models for the FMCE.

Simulations utilizing simple SLD and ULD models, Eqs. 3.1 on page 58 and 3.3 on page 60, constructed distributions of statistics of LD as functions of corresponding LD severity. The statistics of LD used in this study were Pearson's χ^2, G^2, and the tetrachoric correlation. Cutoff values for these statistics of LD were proposed based on the simulation results. These cutoffs were used to identify item pairs where the measured LD was unlikely to result entirely from ULD alone. It was then inferred that some level of SLD must link the items to account for all the LD present. Thus,

the item pairs flagged using this methodology should reveal the effect that chaining items has on the FCI, FMCE, and other assessments.

A supplementary methodology utilized t-testing to compare the statistics of LD for the ULD simulations to those measured for the FCI and FMCE. If the statistics of LD for an item pair were found to be significantly larger that those generated by the ULD simulations, then the item pair was said to likely be linked by SLD.

It was found that the FCI only had two item pairs with LD that could not be explained solely by ULD. This implies that the chaining of items on the FCI is not likely affecting student responses. This result disagrees with previous literature, see Ref. [93]. Further, the results presented in this study imply that the interactions between the FCI and students can likely be modeled using a multi-trait model with minimal errors being introduced in the parameter estimations.

In comparison, the FMCE was found to likely have many SLD linked item pairs. This implies that the FMCE will need to be either graded in a special manner or modified to correct for the detected SLD. Suggestions of how to modify the FMCE were presented in Sec. 3.5 on page 79. Any grading model proposed for the FMCE will need to be studied in detail before it is used in practice/research. Until these studies have been performed, it is recommended that the FMCE not be used, as scores are likely inaccurate. This also holds true for reporting Hake and Normalized student gains on the FMCE.

3.8 Future Direction for Physics Education Research

The results of this study suggest the need for more robust assessments within the field of PER. Before an assessment is used in practice/research, it should undergo *significant* psychometric analysis to ensure it is probing what was intended, in an accurate and fair manner. To enable proper assessments of curricula interventions

(like tutorials, labs, new active learning tools, etc.), the field needs to invest in well-designed, statistically-understood single-conception assessments. Assessments of this nature preserve local independence and can properly assign a single-number score to students and enable accurate differentiation between multiple instruction methods.

This article serves as a cautionary warning to all researchers and instructors who currently use the FMCE. Although it is possible that the simple models used in this study may not fully represent the behavior of LD it is recommended that the future use of the FMCE be paused until more investigations of the instrument are completed. In the meantime, the FCI should be favored over the FMCE for use. Until the potential effects of SLD on multivariate models are better understood, any assessment which likely contains SLD should not be used. That is, if future research into the effects of SLD on student responses finds that SLD significantly impacts the results of educational measures, then it is imperative all previous studies which used the FMCE reconfirm their results.

3.9 Acknowledgments

The authors would like to thank Physport for allowing us to use their data. This project was funded by the Montana State University Physics Department.

CHAPTER FOUR

CONFIRMATORY FACTOR ANALYSIS APPLIED TO THE FORCE CONCEPT INVENTORY

Contribution of Authors and Co-Authors

Manuscript in Chapter 3

Author: Philip Eaton

Contributions: Helped develop and implement the design of this study. Developed code necessary for the project and gave the initial interpretations of the results. Wrote the first draft of the article.

Co-Author: Dr. Shannon Willoughby

Contributions: Helped develop and design this study and oversaw the completion of the project. Provided feedback on analysis and comments on drafts of the manuscript.

Manuscript Information

Philip Eaton, Shannon Willoughby

Physical Review Physics Education Research

Status of Manuscript:

_____ Prepared for submission to a peer-reviewed journal
_____ Officially submitted to a peer-reviewed journal
_____ Accepted by a peer-reviewed journal
__x__ Published in a peer-reviewed journal

Published by the American Physical Society

Published on April 19, 2018, in volume 14, 010124

DOI: 10.1103/PhysRevPhysEducRes.14.010124

Abstract

In 1995, Huffman and Heller used exploratory factor analysis to draw into question the factors of the Force Concept Inventory (FCI). Since then several papers have been published examining the factors of the FCI on larger sets of student responses and understandable factors were extracted as a result. However, none of these proposed factor models have been verified to not be unique to their original sample through the use of independent sets of data. This paper seeks to confirm the factor models proposed by Scott et al. in 2012, and Hestenes et al. in 1992, as well as another expert model proposed within this study through the use of confirmatory factor analysis (CFA) and a sample of 20 822 postinstruction student responses to the FCI. Upon application of CFA using the full sample, all three models were found to fit the data with acceptable global fit statistics. However, when CFA was performed using these models on smaller sample sizes the models proposed by Scott et al. and Eaton and Willoughby were found to be far more stable than the model proposed by Hestenes et al. The goodness of fit of these models to the data suggests that the FCI can be scored on factors that are not unique to a single class. These scores could then be used to comment on how instruction methods effect the performance of students along a single factor and more in-depth analyses of curriculum changes may be possible as a result.

4.1 Introduction

Expertly constructed assessments are used in classrooms is to measure conceptual changes of the students or class compared to other students and classes. As a result, it is of great importance to understand what the assessments are actually measuring. For instance, from an expert's point of view, the Force Concept Inventory [38] is a test of Newton's Laws and the kinematics related to those physical laws. Issues related to that the FCI actually measures have been raised and discussed previously [34, 36, 42].

Recent thrusts of research have investigated this topic using techniques such as exploratory factor analysis (EFA) [83,85,86] and multi-trait item response theory [82]. However, none of these papers have confirmed that either the creators factors nor the factors that they found in their data exists in *other students responses*. In a series

of papers, Huffman and Heller [34, 42] claimed that the factors as laid out by the creators in Ref. [38] do not exists in student responses and that the FCI does not measure what has been claimed. In contrast, this paper uses confirmatory factor analysis (opposed to exploratory) to supply evidence that the measurement model proposed in Ref. [38] fits student data in a satisfactory way.

A measurement model, otherwise known as a factor structure, is a description of how items (e.g., questions of an assessment) load onto an associated factor. Exploratory factor analysis is a tool that attempts to extract the best model for the data that groups the correlations among the student responses. EFA does not attempt to confirm or deny the presence of prespecified models. Since most students think about Newtonian mechanics in a mixture of novicelike and expertlike ways, it is not surprising that EFA does not return a completely expertlike measurement model. Because of this, confirmatory factor analysis (CFA) should be used – and not EFA – to test if the models as suggested in Refs. [38] or [85] actually describe the correlations amongst the questions on the FCI.

Using CFA on a sample of 20 822 student responses, this paper offers evidence that the expertlike model proposed by the creators of the FCI actually models the responses of students in a satisfactory way. Further, the EFA driven model found by Scott et al. [85] was also tested against this set of data. The data used in Ref. [83] for primary misconceptions students had postinstruction. The potential fit of Scott et al.s model onto this alternate data set could lend a suggestion for the primary misconceptions held by postinstruction students in general. Lastly, another expertlike model (created by Eaton and Willoughby) was to be compared to the model proposed in Ref. [38] to see which one better describes student responses. The research questions this paper seeks to answer are how well do the models tested in this study fit a larger sample, and smaller subsamples, and through the use of CFA,

what can be concluded about the factor structure of the FCI and the misconceptions of the students as a result?

In Sec. 4.2 a brief explanation of how the data were obtained will be detailed, followed by an explanation of confirmatory factor analysis, the meaning of model-fit statistics, and a description of how the stability of the models were tested in this paper. The model specifications can be found in Sec. **??** on page ??. The results of the application of the CFA and the stability analysis are presented and discussed in Sec. 4.4 on page 109. The conclusion of the paper follows in Sec. 4.5 on page 120.

4.2 Methodology

4.2.1 Data Collection

The data used in this study came from the PhysPort database [64]. With IRB approval, 22 028 postinstruction student responses for the 1995 FCI were obtained. The students that make up this data set come from both algebra and calculus-based classes, and from potentially all levels of active learning classes, from traditional lecture to flipped classes. The data submitted to PhysPort is self-submitted but has many checks that it must go through before it is admitted into the actual database; details on this process can be found in Ref. [64]. After getting the data, it was cleaned by removing all surveys with any blank responses and further by removing any submissions that were all As, Bs, , Es. After this cleaning process the final data set contained 20 822 student responses.

4.2.2 Explanation of CFA

Confirmatory factor analysis is a subdiscipline of a larger latent variable analysis theory known as structural equation modeling. The purpose of CFA is to confirm that a model proposed by a researcher fits, or sufficiently describes the correlational

groupings of items within a given data sample. This is fundamentally different than exploratory factor analysis, the purpose of which is identifying the measurement model that best describes a specific set of data. As an example, suppose two different sections of an introductory physics course with the same lecturer, teaching style, semester, etc., were given the FCI. When EFA is applied to the response data for these sections, slightly different factors could be found. The differences in the factor structures between these hypothetical sections would be expected to be small; however, significant differences are not impossible. This can happen because EFA is entirely data driven in a way that CFA is not. CFA, however, could be used to compare the fit of the factor model generated by one of the classes onto the other class in an effort to confirm that the correlation structure of the classes is similar. This kind of comparison is not possible with EFA. So, when attempting to verify whether or not a factor structure is present within a set of data EFA is not capable of supplying an answer whereas CFA can provide verification.

EFA and CFA are commonly used tools in psychology research, as well as in economy, sociology, etc. [11]. When developing an assessment, EFA can be used to get a structure for the assessment when no guesses about the factor structure can be made based on expert opinion. This structure found with EFA is used as a model (perhaps with minor modifications) and is tested for validation using CFA on a separate set of data. EFA does not need to be used if the structure of the instrument can be inferred by an expert or is built in, as was attempted by the creators of the FCI.

CFA can be broken up into a four-stage process: (i) model development, (ii) estimation of the models free parameters, (iii) calculation of the model-fit statistics, and (iv) model refinement. Because of the current absence of papers dealing with the application of CFA on conceptual assessments, the steps of CFA will be discussed in

detail here; an in-depth discussion can be found in Ref. [11] and many other textbooks on the subject. If the reader is aware of CFA and how it is done, they can move beyond the remainder of this section as well as the description of the fit statistics in Sec. 4.2.3 on page 97.

The specifications of the measurement model can come from a number of sources. The two most common sources of model specification are a previous application of EFA on another set of data, or theoretical or expert motivation. In both cases the model is a prescription of the number of latent variables (otherwise known as factors), a specification of how latent variables are measured by the items that make up the assessment (item-factor loadings), correlations amongst errors in the residuals, and other considerations. The residual correlation matrix, often called the residuals, is the difference between the true sample correlation matrix and the model generated correlation matrix. For a model with no correlated errors identified, the only parameters that are freely estimated are the loadings of the questions onto their respective factors and the covariance matrix between the latent variables. All other parameters are set to zero in an effort to make the model as parsimonious as possible.

Once the model has been specified, meaning all of the parameters that need to be estimated have been identified, parameter estimation techniques can be used to estimate or calculate the following values: the loading values of the items onto a factor, the reproduced covariance (or correlation) matrices for the items and latent variables, and the residuals. There are a number of standard parameter estimation techniques that are used in practice, such as maximum likelihood, weighted least squares, and Bayesian estimation techniques.

Model fit statistics can be calculated once the model parameters have been estimated. These model fit statistics allow for one to judge the goodness-of-model fit, and they also allow for a comparison of how different models fit the data. A discussion

of the specific fit statistics used in this study can be found after this overview of CFA in Sec. 4.2.3 on the following page.

Models can be refined to improve fit onto a set of data through the use of modification indices and the residuals. Modification indices are a measure of approximately how much the χ^2 of the models fit will decrease if the parameter identified were allowed to be freely estimated within the model, rather than set to zero. Modification indices can suggest two kinds of changes to models being fit: (i) include correlations between the items, or more specifically between the sources of error for the two items, and (ii) change the loading of questions onto different factors.

The first suggestion of the modification indices can arise when two questions are very similar in content and could cause a student to get both questions wrong for related reasons. This correlation can be encapsulated within the model as an added residual correlation parameter. The second suggestion emerges when questions also correlate well with another group of questions other than the factor to which it is currently assigned (i.e., a suggested cross-loading). Within this study, these kinds of suggestions were not heeded in an effort to keep the factor structures unchanged. This is consistent with the goals of this study, in which we seek to validate specific factor structures; altering the loadings of questions onto factors would change the structures and thus could invalidate the conclusions being made.

Further use of the residuals can guide an expert in making modifications to models by identifying correlations that are poorly estimated and include inserting those correlation parameters directly into the model. A detailed description of these statistics, and their uses, can be found in Ref. [11].

With guidance from the modification indices and an experts decision, modifications can be made to models. Following these modifications, the model parameters can be estimated, fit statistics calculated, and modification indices and residuals analyzed

again in an effort to produce a better fitting model. This iterative process can be repeated until a desired model fit has been achieved. In this study, since the sample was made up of post-test responses, novicelike suggestions from the modification indices occasionally were the most favored change to the models. These kinds of modifications were avoided in favor for expertlike suggestions to get a better fit to the data while retaining the expertlike nature of a model. For example, when questions 6 and 7 on the FCI are considered, it can be seen that they ask about related concepts. In both questions the students are asked to identify the path a ball will take after losing either a normal force or a tension force that was causing the ball to travel along a circular path. From an experts perspective it should be expected that if a student gets one of these questions right then they will likely get the other one correct as well, and vice versa. Therefore, we added a residual correlation between questions 6 and 7 in each of the three models.

For this study the open-source R software lavaan [81] was used to help in the estimation of the model parameters using maximum likelihood estimation, calculation of the standard errors of approximation, modification indices, model-fit statistics, and many other useful statistics. In some software packages the covariance between the latent variables is not freely estimated, leaving only the variances to be estimated; lavaans default in this regard is to freely estimate the covariance between all of the latent variables unless otherwise specified.

4.2.3 Model Fit Statistics

Fit statistics used in CFA can be broken up into three categories: absolute fit, parsimony correction, and comparative fit [11]. The statistics that are generally calculated in CFA do not necessarily uniquely fit into one of these categories, but each category can be better described by one statistic over another. It should be

noted that there is debate on what is considered a good model fit for all of the fit statistics presented here; for some papers that have helped guide the debate, see Refs. [12, 13, 41]. The categories of the fit statistics and the statistics themselves are briefly described below.

4.2.3.1 Absolute Fit Absolute fit indices describe how well the model fits the data in an overall sense. This fit does not take into account the fit of the chosen model compared to another model, but is based only on how well the chosen model is able to recreate the correlation matrix of the data. For example, the χ^2 statistic is the difference of the natural logarithms of the determinants of the observed and model generated variance-covariance matrices multiplied by the number of responses minus 1 $[\chi^2 = (\ln|\boldsymbol{S}| - \ln|\boldsymbol{\Sigma}|)(N-1)]$. This is a measure of absolute fit of a single model since it makes no reference to another model when it is calculated. The standardized root mean square residual (SRMSR) is another fit statistics that describes the absolute fit of the model to the data. The SRMSR can be calculated using

$$\text{SRMSR} = \sqrt{\frac{1}{a}\sum_{\substack{i=1\\j<i}}^{n} r_{ij}}$$

where a is the number of elements on and below the diagonal of the correlation matrix, r_{ij} are the elements of the residual correlation matrix, n is the number of items, and the summation is on and below the diagonal of the residual matrix. The SRMSR statistic has values between 0.0 and 1.0, with 0.0 being a perfect model fit and farther away from 0.0 indicating a poorer model fit. For the SRMSR, values close to 0.08 and below are considered in line with good model fit.

4.2.3.2 Parsimony Correction Statistics that fall into this category are different from others in the sense that they introduce penalties for a model having poor

parsimony. A model has poor parsimony, or is not parsimonious, if it contains more freely estimated parameters than needed to achieve good model fit. For example, two models could fit a set of data with the same absolute fit statistics, but one model may be more parsimonious than the other. So, there needs be a way to differentiate between these two models so that the better of the two models can be identified based on fit statistics alone. The parsimony correction index can be used to select the preferred model from the ones with the same, or similar, absolute fit statistics, thereby meeting the goal of using factor analysis to find the most parsimonious model which fits the data. Some very common indices that are used for this category are the root mean square error of approximation (RMSEA), the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). Each of these three statistics uniquely introduces a penalty for a model being nonparsimonious. For example, to calculate the RMSEA the following can be used

$$\text{RMSEA} = \sqrt{\frac{\chi^2_T - df_T}{N df_T}}$$

where df_T is the degree of freedom, χ^2 is the chi-squared statistic of the model being tested, and N is the number of observations or students. The values for the RMSEA begins at 0 and is unbounded above. An acceptable value for this statistic are values close to 0.06 and below, where a value of 0.0 is said to be a perfect model fit.

The other two statistics that take model parsimony into account are the AIC and the BIC, which can be calculated with

$$\text{AIC} = 2b_T - 2\ln(L_T)$$

$$\text{BIC} = b_T \ln(N) - 2\ln(L_T)$$

where b_T is the number of freely estimated parameters, N is the number of observations in the sample, and L_T is the likelihood of the model being tested. As can be seen from their equations, the AIC and the BIC are very similar statistics in that they both reward goodness of fit through the likelihood and penalize for increasing the number of estimated parameters. Thus, smaller AIC and BIC values are indicative of the preferred model in terms of comparison to fit and parsimony. This enables the comparison of models which have differing numbers of factors and items, as is the case in this study. When comparing multiple models to each other, the one with the smallest AIC and BIC values is taken to be the preferred model.

4.2.3.3 Comparative Fit The last category of fit indices are ones that compare the fit of the model being tested to a baseline model. The baseline model takes the covariance between all of the items to be zero, and the variances are freely estimated. As one would expect, since the model being tested is being compared to one that makes no assumption about the relationship between the items, the comparative fit indices often look far more favorable than other fit indices presented. However, in comparative studies some of these indices are found to be some of the best behaved amongst all of the indices presented [11]. The two best behaved statistics are the comparative fit index (CFI) and the Tucker-Lewis index (TLI). These can be found using

$$\text{CFI} = 1 - \frac{\max[\chi^2_T - df_T, 0]}{\max[\chi^2_T - df_T, \chi^2_B - df_B, 0]}$$

$$\text{TLI} = \frac{\chi^2_B - \frac{\chi^2_T}{df_T} df_B}{\chi^2_B - df_B}$$

where χ^2_T and χ^2_B is the model being tested and the baseline models χ^2 value,

respectively, and df_T and df_B are the degrees of freedom of the test model and the baseline model, respectively. The CFI values range from 0 to 1, with 0 indicating no fit, and 1 indicating a good fit for the model compared to the baseline. The TLI calculation can yield values outside of the 0 to 1 range; values less than 0 are rounded up to 0 and values greater than 1 are generally rounded down to 1 [11]. This means the TLI can be interpreted in the same way as the CFI. Accepted fit values for both of these statistics are around 0.90 up to the maximum values of 1 for each statistic. Some sources state that roughly 0.95 and above is indicative of a good model fit, however, there is still debate over what dictates a good model fit, and there is no strict agreement as of yet [11].

4.2.3.4 Local Strain The statistics discussed above describe the models fit to the data in a global sense, meaning they do not look at the residuals individually but all at the same time. Sometimes all of the residuals but one will be within acceptable bounds. These locations of misfit are referred to as a local strain within the model-data fit. Local strain can be found by visual inspection of the modification indices and the residuals, and can be reduced by including residual correlation parameters within the model specifications. It is important to note that the size of the residuals depends on the sample size of the data since the size of the standard error decreases with larger sample sizes. Some methodologists recommend using larger cutoff values for the max allowed residual error as a result [11]. Since the total sample size used for part of this study contained more than 20 000 student responses, the typical bounds for acceptable residuals may be too constraining. When the conventional cutoffs for local strain were enforced for these models, it resulted in models with perfect model fit, and an abundance of residual correlations being added to the models (¿20 in total for all three models). As a result of this, and in an effort to keep the models as

parsimonious as possible, the demand for no local strain was not enforced.

Additionally, two of the models being examined are expertlike models, making the complete removal of local strains difficult. Since the sample of students in this data set is not totally expertlike, the expectation that they will fit these models with absolutely no local strain is unreasonable. This was tested for each of the three models by introducing residual correlations into the models until a fit on half of the total student sample ($\approx$ 10 000 students) that contained no local strain was found. These no strain models, which actually had perfect fits to the data according to most of the fit statistics, were then fitted to the other half of the student data. After inspecting the resulting residuals, it was found that local strain reappeared. Thus, the local strain was found to depend on the composition of the class being considered. Upon investigation of the local strains that developed in the new fit, they were all found to be linked to non-expert-like correlations. As a result, some of the local strains within these models were a result of the unique misfits that manifested from the novicelike nature of the students under investigation. A future study that attempts to construct a model that alleviates all local strain or one that looks at the information the local strains about the students is recommended. This local strain issue does not detract from the conclusions made in this paper about the goodness of fit for the models tested since the four fit statistics used were found to all be within acceptable bounds.

4.2.4 Random Class Generation

For the first part of this study the models were fit to the entire 20 822 student sample. The fit statistics for the models are presented in the results section in Table 4.4. Another goal of this study was to test whether these models could consistently fit smaller sample sizes. The fit, or misfit, of the models when the number of the students in the classes were made smaller will be referred to as the stability of the

model within this study.

A model may fit a large set of data, but as the number of students in the sample decreases the individual misconceptions of each student become more prominent within the correlation structure. Because of the desire to retain the expertlike nature of the models being examined the fit of smaller sample sizes is not guaranteed. If an instructor wanted to investigate how well their class matched an expertlike model they may be unable to use one, or both, of the expert models presented due to the potential instability of the models.

To test the stability of the models, sample classes comprised of 4000, 2000, and 1000 students were uniformly drawn from the ranked 20 822 sample population. This results in smaller samples that have similar means and standard deviations as the sample population they were pulled from. For each of these class sizes 2000 classes were drawn with no duplications in classes, meaning no two classes had the exact same students. The global fit statistics were calculated for each of these classes and the means and standard deviations were calculated for each of the fit statistics. Using the rate of misfit, the stability of each of the models can be determined, with a larger misfit rate indicating a less stable model.

4.3 Measurement model specifications

This study focused on testing three models. The development for two of the models is left to the papers in which they were created. One of these models was found by Scott et al. [85] through the use of EFA on a sample of 2109 postinstruction student responses. This model is called SSG5 (Scott-Schumayer-Gray, 5 factors). The measurement model for SSG5 can be found in Table 4.1 on the next page. Within Tables 4.1 on the following page, 4.2 on page 105, and 4.3 on page 106 the numbers in each column represent the question numbers from the 1995 FCI, the columns are

the factors that the models used, and the double-headed arrows and the $\sim$ indicate correlations that were included in the model.

Table 4.1: This is the factor model found by Scott *et al.* [85] for 2150 students post-instruction. The factor subjects are as follows: Factor 1 = Identification of Forces, Factor 2 = Newton's first law with zero force, Factor 3 = Newton's second law and kinematics, Factor 4 = Newton's first law with canceling forces, and Factor 5 = Newton's third law. The added residual correlations were due to suggested modification indices and from expert consideration of the questions themselves. The double headed arrow and the $\sim$ symbol are being used to represent that a correlation is being estimated between for the two questions on either side.

Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
5	6	19	16	4
11	7	20	17	15
13	8	21	25	28
18	10	22		
30	12	23		
	16	27		
	24			
	29			

Added residual correlations:

5$\sim$18	6$\sim$7	19$\sim$20	8$\sim$23	4$\sim$15
29$\sim$30	10$\sim$24	21$\sim$22	23$\sim$24	

Table 4.2: This is the factor model suggested by Hestenes *et al.* in [38] with modifications made due to questions not fitting with the data. The added residual correlations were due to suggested modification indices and from expert consideration of the questions themselves. The double headed arrow and the $\sim$ symbol are being used to represent that a correlation is being estimated between for the two questions on either side.

Kin.	2nd Law	1st Law	3rd Law	Forces	Superpos.
12	9	6	4	1	17
14	22	7	15	2	25
19	26	8	16	3	26
20	27	10	28	5	
21		23		11	
		24		13	
				18	
				25	
				30	

Added residual correlations:

19$\sim$20	8$\sim$9	6$\sim$7	15$\sim$16	1$\sim$2
21$\sim$22	8$\sim$23	10$\sim$24	23$\sim$24	5$\sim$18

Table 4.3: This is the factor model was developed by Eaton and Willoughby. The added residual correlations were due to suggested modification indices and from expert consideration of the questions themselves. The double headed arrow and the $\sim$ symbol are being used to represent that a correlation is being estimated between for the two questions on either side.

1st Law Kin.	2nd Law Kin.	3rd Law	Force Ident.	Mixed
6	9	4	5	17
7	12	15	11	25
8	14	16	13	26
10	19	28	18	
20	21		30	
23	22			
24	27			
Added residual correlations:				
6$\sim$7	10$\sim$24	21$\sim$22	5$\sim$18	17$\sim$25
8$\sim$23	23$\sim$24	19$\sim$20		

The other two models considered in this study were developed through expert considerations of the questions on the FCI. In Ref. [38] the creators of the FCI proposed a measurement model for the questions on the assessment. This model left alone was found to have many cross loadings and was reduced through the use of model fit statistics and modification indices to make the model more parsimonious. This process resulted in the removal of question 29 from the model due to poor performance. Table 4.2 on the previous page shows the measurement model that came as a result of reducing the original model in Ref. [38] This model is called HWS6 (Hestenes-Wells-Swackhamer, 6 factors).

The (Eaton-Willoughby, 5 factors) EW5 model breaks the questions up into a mixture of the factors identified in Refs. [38, 85] in an effort to create an expertlike model that is capable of fitting smaller sample sizes, which HWS6 had a hard time doing (as discussed in the results section 4.4.2 on page 113). Instead of treating Newtons first and second laws and kinematics as completely different latent variables, they were combined, resulting in the following two factors: Newtons first law with kinematics and Newtons second law with kinematics. Thus, the EW5 model uses the following factors: each of Newtons three laws and their associated kinematics, force identification, and mixed concepts. This model can be found in Table 4.3 on the previous page. Discussion of these factors follows.

The first two factors of the EW5 model are Newtons first and second laws plus the kinematics that result from these laws. In these factors kinematics pertains to path identification and describing how the speed of an object changes for systems that have zero and nonzero net forces, respectively. The questions that were placed into these factors can be found in Table 4.3 on the preceding page. When comparing the Newtons first law factors between the expert models they can be seen to be the same with the exception of one question, question 20. In HWS6 question 20 is put into the kinematics factor of the model, however since EW5 combines Newtons laws and their associated kinematics, question 20 appears in a different factor when the two models are compared. Similarly, questions 12, 14, 19, and 21 (all of the remaining question on HWS5s kinematics factor) moved to the factor in EW5 that combined Newtons second law and its associated kinematics.

The factors classified as Newtons third law in each of the models all have a set of core questions (4, 15, and 28). Question 16 does not appear in the SSG5 model because it did not load in the original EFA analysis done Scott et al. In fact, this question was found by Scott et al. to probe both Newtons first law and not the third

law. In this question, a car pushes a truck while coasting at a constant speed, and from an experts point of view is probing Newtons third law. This question was found by Scott et al. to challenge student understanding of Newtons third law [85].

The factor in EW5 identified as force identification is shared in SSG5, also called force identification. These questions all appear together in the HWS6 model in the forces factor. These questions have been found to create a strong grouping among student responses, and upon inspection all of these questions can be found to be about identifying the forces acting on objects that are stationary or moving at a constant velocity.

The last factor in the EW5 model, called mixed concepts, appears identically in the HWS6 model as the factor superposition principle. Instead of calling it the same name as HWS5, the name mixed concepts was chosen since these questions deal with multiple concepts simultaneously, and not just with superposition of forces. As an example, question 17 is about an elevator being pulled up at a constant speed by a cable. The question asks how the forces acting on the elevator compare to one another. This requires the students to understand how tension works, create a free-body diagram, and then apply Newtons first law to realize that the net force is equal to zero since the system is not accelerating. The other two questions on this factor, questions 25 and 26, are similar to 17 but require an understanding of kinetic friction at an introductory level.

There were some questions in the EW5 model that were left out: 1, 2, 3, and 29. These questions were left out on this expert model due to consideration of the Scott et al. model. They found that these questions did not fit into the EFA factors in a satisfactory way. Therefore, these questions were removed from the expertlike EW5 model in order to avoid poor fit using smaller student data samples.

The residual correlations applied to each of the models were found using

modification indices from the fits of the models to the random subsamples. These correlations had the largest modification indices, were the most common, and were expertlike for the subsample fits. Correlations were added to the models until a nonexpertlike correlation was the largest suggested correction to the models. The procedure resulted in the addition of 9, 10, and 8 residual correlations for the SSG5, HWS6, and EW5 models respectively. Other correlations could be added to improve the fits of the models, but this was not needed as the resulting fits were all within acceptable ranges.

Because of this models consideration of the results in Scott et. al.s resulting EFA factors, particularly the exclusion of questions 1, 2, 3, and 29, this model could be thought of as a hybrid model between expertlike and novicelike. The removal of those four questions was done in an effort to remove poorly performing questions from the model in an attempt to generate an expert model with better fits to the data. The reorganization of the questions from there was done through expert rationale with the intent of not recreating the HWS6 model. The resulting factors as prescribed by EW5 are reasonable from an experts perspective, and as a result this model will continue to be referred to as an expertlike model.

4.4 Results

This section is broken up into two parts, the fit of the models onto the entire sample set, followed by the fit of the models on randomly drawn subsamples of the full sample. The first section shows that each of the models does a good job fitting the entire sample and the second section shows that some of the models have difficulty fitting smaller sample sizes, and are thus referred to as unstable.

4.4.1 Entire Sample

The fit statistics for each model, with no added residual correlations, when fit to the entire data set can be found in Table 4.4 on page 112. All of the models had acceptable fit statistics with no added residual correlations [CFI $>$ 0.9, TLI $>$ 0.9, SRMR $<$ 0.08, RMSEA (Upper CI) $<$ 0.06]. This suggests that the models adequately place questions onto factors in a manner that agrees with the data. Taking parsimony corrections into account, it can be seen that the HWS6 model performs poorly according to the AIC and BIC statistics.

Of all three models analyzed, SSG5 performed the best with the lowest AIC and BIC values of all the models. The goodness-of-fit for SSG5, as well as the other three models, indicates that classes come out of introductory physics with correlation structures that are similar to each other. Suppose the fit for SSG5 had been poor, that would mean that the model generated through EFA performed by Scott et al. was not the best way to represent the correlational groupings of the questions for this large sample. If this were the case then the stability of this model would be called into question, and it could be inferred that classes after instruction potentially have unique factor structures. This would imply that postinstruction, student responses in different classes would be correlated with different topics compared to students from another class. However, since SSG5 did have a good fit to the data, that appears to not be the case, and classes after instruction seem to have the same topical understanding of the questions, as measured by the FCI.

The SSG5 model having the best fit of all the models is not surprising since it is a non-expertlike model that was derived from another student sample, so it inherently models some of the main misconceptions held by students postinstruction. Whereas, HWS6 and EW5 are expertlike models, and the differences in the fits between these expertlike models and SSG5 may be due to the presence of nonexpert thinking.

Further it can be inferred [83] that the primary non-Newtonian world view that students have after instruction is probably the impetus world view.

Of the expert models tested, the AIC and BIC statistics suggest that EW5 fits the student data better than HWS6, that is the correlational structure of the EW5 model more accurately reflects student responses. This may be because HWS6 is too expertlike, and any presence of novicelike correlations causes the fit to be reduced more for this model than the other expertlike model. Ultimately, all of these models do a satisfactory job at describing the relationships between students responses to the FCI. An obvious problem with this particular analysis is that classes are not generally in the 20 000-student range. Investigations into whether these models do a good job at fitting smaller samples was performed.

The results of the models with their residual correlations in place can be found in Table 4.4 on the following page. As expected, the fit statistics for all of the models improved, either increasing or decreasing where appropriate. These results are supplied so that other models can be compared to the three analyzed in this study.

Table 4.4: Fit statistics for the three models applied to the full sample, $N = 20822$, without and with residual correlations (Res. Cor. in the table).

	# of Factors	# of Res. Cor.	CFI	TLI	SRMR	RMSEA (Upper CI)	AIC	BIC
Without residual correlations:								
SSG5	5	0	0.922	0.911	0.032	0.041 (0.042)	538207	538675
HWS6	6	0	0.911	0.900	0.037	0.040 (0.041)	654026	654622
EW5	5	0	0.915	0.904	0.038	0.042 (0.043)	585590	586082
With some residual correlations:								
SSG5	5	9	0.973	0.968	0.021	0.025 (0.025)	532708	533248
HWS6	6	10	0.949	0.941	0.033	0.031 (0.032)	648887	649554
EW5	5	8	0.955	0.948	0.032	0.031 (0.032)	580544	581100

4.4.2 Subsamples of the entire sample set

The 2000 subsamples of 4000, 2000, and 1000 student classes, respectively, were generated by randomly drawing, without replacement, students from the entire 20 822 sample. The fit statistics were calculated for each of these classes for each model, and the results are presented in Tables 4.5 on page 115, 4.6 on page 117, and 4.7 on page 119 for the SSG5, HWS6, and EW5 models, respectively. We will go through each models performance one by one, after first describing the meaning of misfit in more detail.

A misfit results when the maximum likelihood algorithm converges to a nonacceptable solution. Occasionally errors such as variances for questions will come out greater than 1 or negative, correlation matrices for the latent variables may be nonpositive definite, correlation between questions may be greater than 1 or negative, etc. These kinds of errors indicate a model is ill specified for the data that it is being fit to, otherwise known as a misfit. When the models were being fit to the randomly sampled classes all instances that resulted in a misfit of the model to a class were counted. After all of the classes had been fit to the models the percentage of the classes that misfit the models were calculated. A cutoff percentage of 15% was used as an indication that a model had fundamental issues when trying to fit the smaller sample sizes. As a result, none of the fit statistics were presented for any model that had a misfit rate above 15% due to the instability of the model at that sample size.

The actual value used for the cutoff, 15%, was chosen in the spirit of the most lenient p-value commonly used in hypothesis testing of $p < 0.15$. The p value is the probability of finding the observed data when the null hypothesis is true. Generally, the null hypothesis is that the model being tested is stable, so the 15% cutoff for the misfit rate carries a similar meaning to testing the null hypothesis. Ultimately, this cutoff rate is arbitrary and could be increased or decreased to change the rigor of the

stability testing.

Table 4.5 shows the performance for the SSG5 model. As can be seen there were no misfits for any of the sample sizes (which is not the case for the other models) and the fit statistics were satisfactory to claim good model-data fit. In fact, the fit statistics can be seen to change little from the 4000 student classes to the 1000. This indicates that this model is stable for the smaller sample sizes, and that using this model to fit individual classes may be possible. Since Scott et al. [83] found the most prominent misconceptions contained in their data, an instructor may want to check the fit of this model to their students to verify if they potentially possess similar misconceptions.

Table 4.5: Fit statistics' mean and standard deviations of 2000 samples of size 4000, 2000, and 1000 students for the SSG5 model. None of the smaller samples tested had any misfits, which is an indication that this model is fairly representative of the general population. As the sample sizes get smaller the fit statics get worse, but never get to the point of representing a poor model fit.

Scott-Schumayer-Gray 5 factors – SSG5

Without Residual Correlations:

	Mean	St. Dev.		Mean	St. Dev.		Mean	St. Dev
4000 Students			**2000 Students**			**1000 Students**		
CFI	0.921	0.0043	CFI	0.920	0.0065	CFI	0.918	0.0097
TLI	0.910	0.0049	TLI	0.908	0.0074	TLI	0.906	0.0112
SRMR	0.034	0.0009	SRMR	0.036	0.0013	SRMS	0.039	0.0018
RMSEA	0.041	0.0011	RMSEA	0.042	0.0017	RMSEA	0.042	0.0026
RMSEA Upper CI	0.043	0.0011	RMSEA Upper CI	0.044	0.0017	RMSEA Upper CI	0.046	0.0025
AIC	103440	528	AIC	51737	391	AIC	25878	284
BIC	103812	528	BIC	52067	391	BIC	26168	284

With Residual Correlations:

4000 Students			**2000 Students**			**1000 Students**		
CFI	0.972	0.0024	CFI	0.971	0.0039	CFI	0.969	0.0064
TLI	0.967	0.0029	TLI	0.966	0.0046	TLI	0.963	0.0076
SRMR	0.023	0.0008	SRMR	0.026	0.0012	SRMS	0.031	0.0016
RMSEA	0.025	0.0011	RMSEA	0.025	0.0017	RMSEA	0.026	0.0028
RMSEA Upper CI	0.027	0.0011	RMSEA Upper CI	0.028	0.0017	RMSEA Upper CI	0.026	0.0026
AIC	102372	531	AIC	51217	385	AIC	25620	289
BIC	102800	531	BIC	51598	385	BIC	25953	289

The HWS6 model had a relatively large number of misfits using the smaller sample sizes. In fact, for the samples of 2000 and 1000 students this model misfit the classes more that 15% of the time. However, when the model did not have a misfit it retained a good fit with the data. Because of the large misfit rate this model can be concluded to be unstable for small sample sizes. This instability at smaller sample sizes may be due to the fact that this model is too expert-like, as mentioned previously. As the sample sizes get smaller and smaller, misconceptions of an individual student become more prominent within the correlation matrix. As a result even subtle differences in response patterns from a handful of students could be enough to cause a misfit between this model and the data. This is conjecture, and more analysis should be done in a future study to address the specifics for why this model does so poorly at smaller sample sizes.

Table 4.6: Fit statistics' mean and standard deviations of 2000 samples of size 4000, 2000, and 1000 students for the HWS6 model. Many of the smaller samples tested misfit with the model, which is an indication that this model is not a good representation of smaller samples.

Hestenes-Wells-Swackhamer 6 factors – HWS6						
			Without Residual Correlations:			
	Mean	St. Dev.		Mean St. Dev.		Mean St. Dev
4000 Students	Misfit rate = 5.05%		**2000 Students**		**1000 Students**	
CFI	0.910	0.0039	CFI		CFI	
TLI	0.899	0.0044	TLI		TLI	
SRMR	0.039	0.0010	SRMR		SRMS	
RMSEA	0.040	0.0009	RMSEA	Misfit rate > 15%	RMSEA	Misfit rate > 15%
RMSEA Upper CI	0.042	0.0009	RMSEA Upper CI		RMSEA Upper CI	
AIC	125664	585	AIC		AIC	
BIC	126136	585	BIC		BIC	
			With Residual Correlations:			
4000 Students	Misfit rate = 3.10%		**2000 Students**		**1000 Students**	
CFI	0.950	0.0028	CFI		CFI	
TLI	0.942	0.0033	TLI		TLI	
SRMR	0.034	0.0011	SRMR		SRMS	
RMSEA	0.031	0.0009	RMSEA	Misfit rate > 15%	RMSEA	Misfit rate > 15%
RMSEA Upper CI	0.032	0.0009	RMSEA Upper CI		RMSEA Upper CI	
AIC	124655	580	AIC		AIC	
BIC	125190	580	BIC		BIC	

The results of the other expert model, EW5, can be found in Table 4.7 on the following page. This model appears to be a better representation of how the questions fit together as the misfit rate is drastically lower compared to the HWS6 model. The fit statistics are good for all of the categories and change very little as the class size decreases. Comparing the AIC and BIC for this model and HWS6 for 4000 students in a class, it can be seen that this model is better at describing the data with and without residual correlations in place. Between the expert models, EW5 appears to do a better job in general at fitting the data to latent variables.

Table 4.7: Fit statistics' mean and standard deviations of 2000 samples of size 4000, 2000, and 1000 students for the EW5 model. Some of the smaller samples tested misfit with the model, but the rate of misfits was never above 3% for the model with correlations in place. This suggests that for sample smaller than 1000 students the model without correlation should be fitted to the data first and then correlation can be added after model-data fit has been established. As the sample sizes get smaller the fit statics get worse, but never get to the point of representing a poor model fit.

Eaton-Willoughby 5 factors – EW5

Without Residual Correlations:

	Mean	St. Dev.		Mean	St. Dev.		Mean	St. Dev
4000 Students			**2000 Students**			**1000 Students**		
CFI	0.914	0.0039	CFI	0.913	0.0061	CFI	0.911	0.0094
TLI	0.903	0.0044	TLI	0.902	0.0068	TLI	0.899	0.0106
SRMR	0.040	0.0011	SRMR	0.041	0.0015	SRMS	0.045	0.0022
RMSEA	0.042	0.0010	RMSEA	0.043	0.0015	RMSEA	0.043	0.0024
RMSEA Upper CI	0.044	0.0010	RMSEA Upper CI	0.045	0.0016	RMSEA Upper CI	0.047	0.0023
AIC	112521	516	AIC	56271	395	AIC	28170	284
BIC	112912	516	BIC	56618	395	BIC	28474	284

With Residual Correlations:

4000 Students			**2000 Students**	Misfit rate = 0.30%		**1000 Students**	Misfit rate = 2.85%	
CFI	0.954	0.0028	CFI	0.953	0.0046	CFI	0.951	0.0069
TLI	0.947	0.0032	TLI	0.946	0.0053	TLI	0.943	0.0080
SRMR	0.034	0.0011	SRMR	0.036	0.0017	SRMS	0.039	0.0022
RMSEA	0.031	0.0010	RMSEA	0.032	0.0016	RMSEA	0.033	0.0024
RMSEA Upper CI	0.033	0.0010	RMSEA Upper CI	0.034	0.0016	RMSEA Upper CI	0.036	0.0023
AIC	111564	539	AIC	55788	409	AIC	27925	297
BIC	112005	539	BIC	56180	409	BIC	28269	297

For the EW5 model there were a few cases of misfit at the lower sample sizes with residual correlations in place. This makes sense since the residual correlations that were included were only expertlike. This means the model with the residual correlations in place can be considered to be more expertlike compared to the measurement model without correlations in place. As a result, classes with more novicelike correlations will potentially have a harder time fitting this model and thus the misfit rate increases. This can be seen to be the case in Table 4.7 for the 2000 and 1000 sample sizes comparing the misfit rates with and without the residual correlations. As a result, if an instructor wanted to try to fit this model to their own classs results they should consider starting with the correlation free model to initially see how well their class fits the expertlike EW5 model.

4.5 Conclusions

Confirmatory factor analysis was applied to three models using a set of data with 20 822 postinstruction student responses to the Force Concept Inventory. Of the models, one was found through the use of exploratory factor analysis applied to 2109 students by Scott et al. [85]. The other two models were expert created models without the use of EFA, described in the model specification section. All of these models were found to fit the full sample with satisfactory fit values. This means that all three of these models could be used to describe the correlations between students responses to the FCI after instruction.

Further analysis of the SSG5 data by Scott et al. [83] revealed that the impetus world view was the primary misconception held by the students whose response data generated the SSG5 model. Because of the good model fit of SSG5 with this larger data set it could suggest that the impetus world view is the chief issue students still have after instruction.

The fact that the HWS6 model has an acceptable fit with the full sample ($N = 20\,822$) suggests that this factor structure accurately represents postinstruction student responses on the FCI. Through the use of CFA (instead of EFA) this study finds that it agrees with Halloun and Hestenes in that the factor structure presented in Ref. [38] is an acceptable way to categorize the questions of the FCI.

This conclusion disagrees with the conclusions of Huffman and Heller presented and discussed in Refs. [34, 42]. The disagreement between these studies could primarily be a result of the statistical tools chosen to answer the research question. EFA as a tool is not capable of confirming or denying the existence of a given factor structure within a set of data. This is not to say that the validity of a model cannot be inferred through the use of EFA on alternate sets of data, which is a common application. That no consistent model was found through the application of EFA on different sets of data may imply that the factors of the FCI are sensitive to the sample being analyzed. This study demonstrates that the FCI does measure what it was design to measure as described by Hestenes et al. from a factor perspective. Given the results seen herein, we hope to end the air of caution carried by researchers about the factor structure of the FCI and to end the debate that began in 1995.

Because these expert factor models have been confirmed for the FCI, instructors can now look at graded chunks of the FCI. For instance, using the EW5 model, the FCI can be thought of as testing the 5 factors indicated in the model specifications. Thinking specifically about Newtons third law, an instructor can inspect how a student (or the whole class) answered questions 4, 15, 16, and 28 and determine the extent to which Newtons third law is understood after instruction. This is a simple example of what can be done, but the power this gives instructors for targeting concepts their students are struggling with could help in assessing instruction methods for specific sections of the material.

After examining the fits of these models to smaller samples constructed from the larger data set, it was found that the SSG5 and the EW5 models did a good job fitting with little or no misfits. HWS6, however, had a hard time fitting these smaller samples. This seems to indicate that when students have conceptual issues with Newtons first or second laws they may also have difficulties with their associated kinematics. This is opposed to viewing Newtons laws and kinematics as being entirely separate factors and thus give the impressions that one could grasp one topic without fully understanding the other.

4.6 Future Work

Further research into why the HWS6 model has a hard time fitting the smaller sample sizes is suggested. Also, investigating how these models fit data that is only from an algebra- or calculus-based class and a comparison of these results is currently being pursued. Other affects, like gender or different teaching styles, on the fit of these models is also being considered.

An investigation into what the suggested modification indices can reveal about classwide postinstruction misconceptions in an exploratory or confirmatory factor analysis style of analysis is being investigated as well.

4.7 Acknowledgments

The authors would like to thank Keith Johnson and Barrett Frank for reading over the manuscript, offering insightful suggestions, and the many interesting discussions related to this project. This project was funded by the Montana State University Physics Department.

CHAPTER FIVE

IDENTIFYING A PREINSTRUCTION TO POSTINSTRUCTION MODEL FOR THE FORCE CONCEPT INVENTORY WITHIN A MULTITRAIT ITEM RESPONSE THEORY FRAMEWORK

Contribution of Authors and Co-Authors

Manuscript in Chapter 4

Author: Philip Eaton

Contributions: Helped develop and implement the design of this study. Developed code necessary for the project and gave the initial interpretations of the results. Wrote the first draft of the article.

Co-Author: Dr. Shannon Willoughby

Contributions: Helped develop and design this study and oversaw the completion of the project. Provided feedback on analysis and comments on drafts of the manuscript.

Manuscript Information

Philip Eaton, Shannon Willoughby

Physical Review Physics Education Research

Status of Manuscript:

_____ Prepared for submission to a peer-reviewed journal
_____ Officially submitted to a peer-reviewed journal
_____ Accepted by a peer-reviewed journal
__x__ Published in a peer-reviewed journal

Published by the American Physical Society

Published on 28 January, 2020

DOI: 10.1103/PhysRevPhysEducRes.16.010106

Abstract

As targeted, single-conception curriculum research becomes more prevalent in physics education research (PER), the need for a more sophisticated statistical understanding of the conceptual surveys used becomes apparent. Previously, the factor structure of the Force Concept Inventory (FCI) was examined using exploratory factor analysis (EFA) and exploratory multitrait item response theory (EMIRT). The resulting exploratory driven factor model and two expert-proposed models were tested using confirmatory factor analysis (CFA) and were found to be adequate representations of the FCI's factor structure for postinstruction student responses. This study compliments the literature by using confirmatory multitrait item response theory (CMIRT) on matched preinstruction to postinstruction student responses ($N = 19\ 745$). It was found that the proposed models did not fit the data within a CMIRT framework. To determine why these models did not fit the data, an EMIRT was performed to find suggested modifications to the proposed models. This resulted in a bi-factor structure for the FCI. After calculating the student ability scores for each trait in the modified factor models, only two traits in each of the models were found to exhibit student gains independent of the general, "Newtonian-ness" trait. These traits measure students' understanding of Newton's third law and their ability to identify forces acting on objects.

5.1 Introduction

Concept inventories have become some of the most widely used tools in physics education research (PER) since their introduction. One of the first successful conceptual inventories created was the Force Concept Inventory (FCI), which was constructed in 1992 and updated in 1995 [38]. Because of the success of this instrument many similar assessments have been created to probe conceptual topics in physics, such as the Force and Motion Conceptual Evaluation [97], the Energy and Momentum Conceptual Survey [87], the Conceptual Survey of Electricity and Magnetism [60], the Brief Electricity and Magnetism Assessment [21], and the Thermal Concept Evaluation [105], among others. Many studies have tested these assessments in an attempt to understand the psychometric properties of these tools, see Refs. [21, 22, 43, 60, 78, 87, 90, 96, 97, 105].

Within PER, conceptual assessments, such as the FCI, are often used to identify effective research-based teaching strategies and curricula. For example, a significant portion of the evidence used to support the utilization of active learning techniques comes from student preinstruction to postinstruction gains on validated conceptual assessments [29, 31, 76]. The assessments used in these types of studies must be well understood to lend validity and strength to the conclusions derived from their statistics.

The psychometric analysis performed thus far on the FCI can be effectively broken up into two types: unidimensional analysis and multidimensional analysis. The unidimensional analyses performed on the FCI include, but are not limited to, item response curves [67, 68], (unidimensional) item response theory (IRT) [75, 99, 104], and classical test theory (CTT) [29, 31, 35, 49, 76, 99]; multidimensional studies include, but are not limited to, exploratory factor analysis (EFA) [83, 85, 86], confirmatory factor analysis (CFA) [26], multitrait (multidimensional) item response theory (MIRT) [82, 93], cluster analysis [91, 92], and network analysis [84, 106]. These studies have been summarized in Table 5.1 on the next page.

Previously, we used CFA to test the factor structure of the FCI [26]. This analysis was driven in part by the "H&H debate" [34, 36, 42], which discussed the validity and stability of the FCI's factor structure. This study found that the original factor structure proposed by the creators of the FCI did not fit student response data from the assessment. However, the creator's model was found to fit the data adequately after adjustments, obtained through CFA modification indexes, were made to the creator-proposed model [26]. The CFA study also investigated two other factor models, one expert proposed and another that was driven by the EFA of the FCI presented by Scott and Schumayer in Ref. [85], and found them both to adequately explain the student response data for the FCI.

Table 5.1: Analyses that have been performed on the FCI. This list is by no means complete.

Unidimensional Analyses	
Item response curves	[67, 68]
Item response theory	[75, 99, 104]
Classical test theory	[29, 31, 35, 49, 76, 99]
Multidimensional Analyses	
Exploratory factor analysis	[83, 85, 86]
Confirmatory Factor Analysis	[26]
Multitrait item response theory	[82, 93]
Cluster analysis	[91, 92]
Network analysis	[84, 106]

This study intends to extend the claims of our previous CFA work by using MIRT. It should be noted that MIRT has been used to study the FCI previously, see Refs. [82] and [93]. Scott and Schumayer [82] used MIRT to verify an exploratory factor model they found using exploratory factor analysis in Ref. [85]. This model is used in this study as one of the models being tested. Stewart *et al.* [93] presented a theoretically driven MIRT model that attempted to model the fragmented thought process of students while taking the FCI. However, only parsimonious fit statistics were reported for this model. To help identify if the model had good fit with the data a standardized root mean square of the residuals (SRMSR) should have, at minimum, been presented. We used Stewart *et al.*'s model and fit it to the data gathered for this study and found that there was an inadequate goodness of fit with the data,

with a SRMSR > 0.1. Although the model presented by Stewart *et al.* has a strong theoretical basis, it does not adequately represent data from the FCI used in this study.

The purpose of this study is twofold: (i) to strengthen the CFA results of Ref. [26] using MIRT to supply a deeper understanding of the complex factor structure of the FCI, and (ii) to identify the concepts the FCI can probe independent of other factors in the factor model using a MIRT framework. This will be done by answering the following research questions:

RQ1: Do the expert-proposed models tested in Ref. [26] adequately explain the data in a confirmatory multitrait item response theory framework?

RQ2: Using exploratory multitrait item response theory as a guide, can the expert-proposed models from Ref. [26] be modified to better fit the data, and what do these modifications suggest about the FCI?

RQ3: Provided an adequate confirmatory model is found, which of the models' factors demonstrate measurable student gain from preinstruction to postinstruction independent of the other factors?

The article is organized as follows: Sec. 5.2 on the following page discusses the characteristics of the data used in this study as well as how the data was prepared for analysis. Next, a brief explanation to motivate the need for a multitrait model is presented in Sec. 5.3 on page 130, followed by the methodology in Sec. 5.4 on page 132. The results of the methodology and a subsequent discussion can be found in Secs. 5.5 on page 147 and 5.6 on page 152. Lastly, Secs. 5.7 on page 155, 5.8 on page 156, and 5.9 on page 156 address the limitations of, future works related to, and the conclusions of this study, respectively.

5.2 Data

The data used in this study were received from PhysPort [64]. PhysPort is a support service that synthesizes physics education research and details effective curricula for instructors. Further, PhysPort collects and manages a plethora of PER assessments and maintains a database of student responses for many of these assessments.

The original PhysPort data used in this study consisted of 22 029 matched preinstruction-to-post instruction student responses for the 1995 version of the FCI. The sample was made up of student responses from a mixture of algebra- and calculus-based introductory physics I courses and a mixture of teaching strategies (e.g., traditional lecturing, flipped classroom, etc.). Statistical results from a dataset of this kind will thus pertain to the more general population of "students in introductory physics I." This means the effects of variables such as "mathematical sophistication of the course" and "teaching strategies" cannot be assessed using this data (among others, like student's gender, student's academic year taking the course, the semester or quarter the course was taken, etc).

Any student responses that contained a blank response were removed from the sample in a listwise deletion, which left 19 745 student responses in the sample. These 19 745 student responses were then split in into two halves using their even and odd identification numbers given in the original PhysPort data. The odd half of the data, 9841 student responses, was used to generate the exploratory models, which were used to suggest possible modifications to the expert models. The even half of the data, 9904 student responses, was used to test the fit of the exploratory model to a different, independent sample to supply evidence for the possible generalizability of the results. Further, both the odd and even halves of the data were used to assess

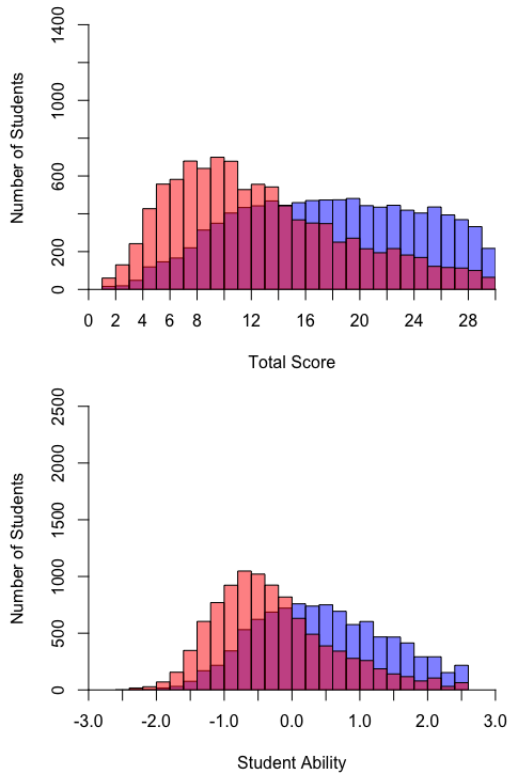
the goodness of fit of the original and modified expert-proposed models. It should be noted that the data used in this study and in Ref. [26] are the same, with the exception of the removal of students who did not complete both the preinstruction and postinstruction FCI.

5.3 Motivation for a multitrait analysis

Unidimensional models inherently assume that the assessment being modeled only measures a single trait, or student ability. Models of this nature enable researchers and instructors to readily compare students' performances to one another and to compare a single student's performance on an assessment from preinstruction to postinstruction through the use of one statistic (i.e., aggregate total score or a unidimensional IRT student ability score).

Assuming an assessment is measuring a single trait may oversimplify, and potentially misrepresent, what students specifically learned from preinstruction to postinstruction. A pre-to-post change in a unidimensional total score will indicate that the students learned something relevant to the assessment, but does not specify which concepts may have been learned. For instance, Fig. 5.1 on the following page shows the FCI scores for 9841 matched preinstruction to postinstruction student responses using the aggregate total score (total number of questions correct) and the student ability scores from a one-dimensional two-parameter logistic model from IRT. These results indicate that student abilities from IRT often do not add more information than the aggregate total score. A linear regression between estimated student abilities and total scores for the FCI found a regression correlation of $r^2 = 0.994$ [104]. This suggests that the total score and student ability on the FCI give effectively the same information. It can be seen that both methods of scoring the assessment indicate that the students did improve over the course of instruction.

Figure 5.1: Histograms of student scores preinstruction (red) and postinstruction (blue). The top plot used the traditional total score for students found by adding up their total number of correct responses, and the bottom plot used the estimated student abilities from a one-dimensional two-parameter logistic model in IRT. Both methods show a flattening to the distribution of scores, and a definite shift to higher scores, which indicates the students performed better on the assessment postinstruction.



However, these differences are blind as to which conceptions the students' specifically learned on the assessment.

Conclusions based only on these models could be misleading depending on the research questions being asked. For example, an assessment could be built to probe Newton's first, second, and third laws independently from one another; an assessment like this would be said to have scales representing the three laws of Newton.

Suppose this assessment was scored using a unidimensional model for preinstruction and postinstruction results to investigate the effectiveness of new curricula which specifically targets Newton's first law. Further, suppose that the students did not effectively learn Newton's First Law, but did have "strong" gains along the Newton's second and third law scales. The results from the unidimensional scoring would be blind to this specific situation, and would conclude that the new curricula did effectively teach Newton's first law, even though it actually did not. Although this example is extreme, it succinctly demonstrates the crucial need for physics education researchers to understand the factors present on the assessments being used in practice.

To properly measure student growth on an assessment that claims to measure multiple conceptions, a multidimensional (or multitrait) model should be used. There are many ways to identify and generate multidimensional models for assessments, like factor analysis and multitrait item response theory (MIRT). We chose to use MIRT since it is a model driven analysis that readily assigns scores to students in the form of an estimated student ability vector. A brief explanation of MIRT and how the models used in this study were created will be discussed in the following section.

5.4 Methodology

The following discussions were kept short for brevity purposes; references cited within each subsection can be used to find more detailed information about the tools and techniques used in this study.

In short, the expert proposed models from Ref. [26] (see Tables 5.2 on the next page and 5.3 on page 135) were tested using confirmatory MIRT (CMIRT). Following this, exploratory MIRT (EMIRT) was preformed on the odd half of the data to find suggestions for how the expert models could be improved if necessary. Subsequently,

Table 5.2: The qualitative interpretations for each of the factors for the bi-factor expert-proposed models used in the CMIRT analysis.

Hestenes, Wells, and Swackhamer 6-Factor Model (HWS6)	
F1:	Newton's first law
F2:	Newton's second law
F3:	Kinematics
F4:	Forces
F5:	Newton's third law
F6:	Superposition of forces
Eaton and Willoughby 5-Factor Model (EW5)	
F1:	Newton's first law with kinematics
F2:	Newton's second law with kinematics
F3:	Identification of forces
F4:	Newton's third law
F5:	Superposition of forces and mixed methods
Scott and Schumayer 5-Factor Model (SS5)	
F1:	Newton's first law with zero netforce
F2:	Newton's second law with kinematics
F3:	Identification of forces
F4:	Newton's third law
F5:	Newton's first law with canceling forces

CMIRT was used again to reexamine the modified expert models to verify good model fit with both the odd and even halves of the data. Once good model fit was

verified, the students' abilities were estimated and plotted on radar plots, discussed in Sec. 5.4.5 on page 143.

5.4.1 Multitrait item response theory

Item response theory postulates that the manner in which a student responds to an item is a consequence of the interaction between the student and item characteristics. MIRT further postulates that a single assessment can probe multiple traits (i.e., latent abilities). As a result, students and items will have different characteristics for each trait in the multitrait model of the assessment. For example, each item's will have different characteristics and students will have different ability scores for each trait in the model. This enables researchers and instructors to determine which traits a student is excelling in and traits a student is in need of assistance.

The characteristics of an item are specified within a mathematical model called an item response function. An example of an item response function is the multitrait 2-parameter logistic (2PL) function [65], which can be written in the following manner:

$$P_{ij}(x_{ij} = 1|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, \delta_j) = \frac{1}{1 + e^{-D(\boldsymbol{\alpha}_j \cdot \boldsymbol{\theta}_i - \delta_j)}}$$

where P_{ij} is the probability that student i, with and ability vector $\boldsymbol{\theta}_i$, will answer item j correctly given the item's discrimination vector and difficulty, $\boldsymbol{\alpha}_j$ and δ_j. The constant D is generally taken to be 1.702 to fix the scale of the logistic model such that $\theta = 1$ approximately corresponds to 1 standard deviation of the sample's student ability scores along a single trait. To fully describe the interaction of student i and item j along trait k one needs to know the student's trait ability θ_{ik}, the item's trait discrimination α_{jk}, and the item's difficulty δ_j. Since items are allowed to cross

load onto multiple traits, each item may have multiple trait discriminations, one for each trait the item loads onto. The interpretation of the item parameters is not straightforward, but becomes easier when considering a single trait (i.e., setting all other trait abilities to zero).

Table 5.3: The original expert-proposed models tested using CFA in Ref. [26].

Hestenes, Wells, and Swackhamer 6-Factor Model (HWS6)						
F1	F2	F3	F4		F5	F6
6	9	12	1	13	4	17
7	22	14	2	18	15	25
8	26	19	3	25	16	26
10	27	20	5	30	28	
23		21	11			
24						

Eaton and Willoughby 5-Factor Model (EW5)						
F1		F2		F3	F4	F5
6	20	9	21	5	4	17
7	23	12	22	11	15	25
8	24	14	27	13	16	26
10		19		18	28	
				30		

Scott and Schumayer 5-Factor Model (SS5)						
F1		F2		F3	F4	F5
6	12	19	22	5	4	16
7	16	20	23	11	15	17
8	24	21	27	13	28	25
10	29			18		
				30		

For instance, the trait discrimination of item j along trait k, α_{jk}, describes the slope of the item response function at a probability of 50% along the axis of trait k. The larger α_{jk} is the more discriminating the trait is for that item, or the better the item is at differentiating between students with abilities above or below the difficulty of the item within that trait. So, students will generally need smaller gains along traits on which items have large discrimination values to significantly improve their odds of answering the question correct, and vice versa. Typically, a single trait will contain a group of items that have large discrimination values compared to the other items on the assessment. These groups of items can be treated in a similar manner as the factors found through exploratory factor analysis [18].

Notice the item difficulty is the same regardless of the trait being considered. A trait difficulty for item j along trait k can be expressed as δ_j/α_{jk}, and can be shown to be the location along that trait's ability axis (all other trait abilities set to zero), where the probability of getting the question correct is 50%. So, the larger the discrimination an item has along a trait, the easier the item appears and vice versa. This is consistent with the idea that the better the item discrimination is along a trait, the less a student needs to improve that trait to significantly improve their odds of getting the item correct.

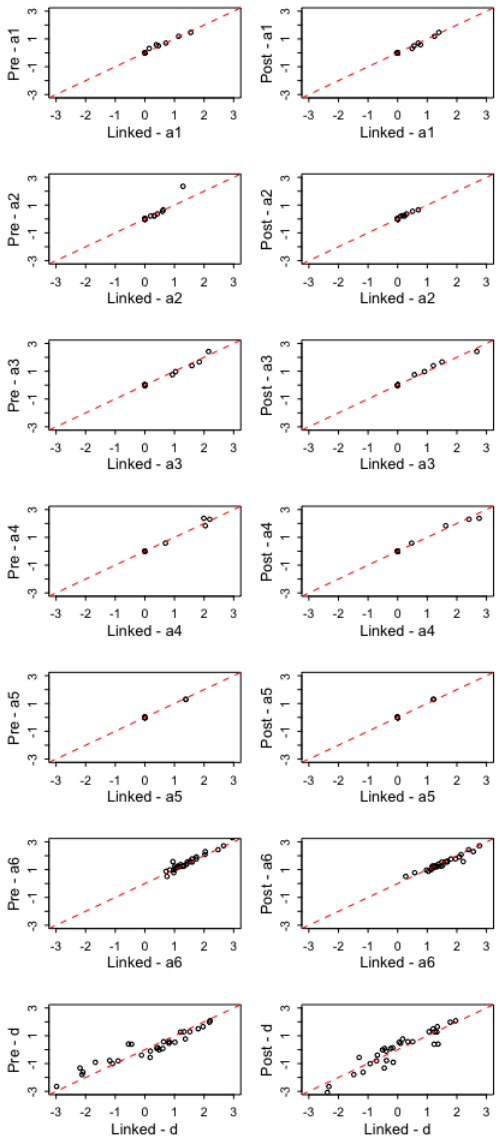
A more detailed discussion of MIRT can be found in the many resources dedicated to the subject. These details, as well as the parameter estimations techniques used to find the item characteristics and student ability scores, are out of the scope of this paper; more technical details can be found in the literature [4, 4, 18, 66, 73, 74, 82, 93]. This study used maximum likelihood parameter estimation techniques using the the open-source software package "mirt" [14] based in the programming language R [77].

5.4.2 Linking multitrait item response theory models

Because of the dependency on the data to set the metric in which the item parameters are estimated, one will need to link the metrics of models when trying to compare results from multiple sets of data. Within a unidimensional IRT model this linking appears in the form of a rescaling and a translation of the student ability metric. This becomes more complicated in a multitrait model where rescaling and translations need to be done along each of the traits, after rotating the ability axes to ensure the models are measuring the same traits. This can be done using a number of different methods, the details of which can be found in Refs. [46, 51–53, 111].

For this study, we decided to use a mean anchored (or mean anchoring) linking procedure. This was done by specifying the groups of questions within the MIRT model through prescribing the factor structure of the assessment, which is known as confirmatory multitrait item response theory. With the factor model in place, the preinstruction and postinstruction samples were used to find the item parameters for the specified model. The resulting item difficulty values were shifted from student centered to item centered by subtracting the mean item difficulty from all of the difficulty parameters for each sample. Finally, the pretest and post-test item parameters were averaged and these mean values were used to estimate the student abilities for both samples; thus the name: mean anchored. Goodness-of-fit statistics for the parameters generated using this linking procedure were found to be acceptable and are reported in the Sec. 5.5 on page 147 when discussing the results of each of the models used in this study.

Figure 5.2: Preinstruction and postinstruction models original estimated item parameters (y axis) plotted versus the combined mean anchored item parameters (x axis). The red, dashed line is the 1-to-1 line, a line of slope 1 and y-intercept 0. The points are, for the most part, well regressed onto the 1-to-1 line, meaning the item parameters are in good agreement with one another. For these plots a1–a6 represent the item discriminations for the 6 traits in the model and d stands for the item difficulty. Specifically, these plots are for the EW5 model, see Table 5.6 on page 149



After a scale linking is performed, it is important to check the quality of the linking. This can be done by plotting the item parameters (i.e., the trait discriminations and difficulties) of the original scale to the linked scale [46]. For example, Fig. 5.2 on the previous page displays the relationship between the preinstruction, postinstruction, and the mean item parameters for the Eaton and Willoughby 5-factor model [26]. From these plots it can be seen that the item parameters are in good agreement with one another, with minor exceptions. Similar plots for the other models were generated, but did not differ much from these results. So, these plots were left out of the article for brevity. These plots, in conjunction with the acceptable goodness-of-fit statistics that result from models treated in this manner, indicates that the utilization of the mean anchoring procedure is acceptable.

5.4.3 Exploratory multitrait item response theory

One way to identify potential factor models that could be used in a CMIRT analysis is to explore the factor structure of the sample using an exploratory analysis. This was done by generating exploratory models ranging from 3 factors up to 7 factors. To ensure the resulting models would be expertlike in nature, postinstruction student responses were used to initially establish each model's factor structure.

The initial exploratory models allowed all of the questions to load onto all of the available factors. From here, questions were systematically removed from individual factors to create a more parsimonious factor structure. This was done by removing the smallest calculated item loading from the model one by one, while monitoring the model's goodness of fit. That is, the loading parameter that corresponded to the smallest loading value in the factor model was removed for a single iteration of this process.

Item loading values explain the amount of variance a factor explains for a

particular item; a factor represents a trait in both factor analysis and MIRT. The closer the magnitude of an item loading is to 1, the more variance of the item is explained by the factor. As a result, a loading value is interpreted as being small if its magnitude is close to zero.

Beginning with item loading values with magnitudes below 0.05 for postinstruction data, items were removed from individual traits one at a time. After an item was removed, the new model was fit to the data and the goodness-of-fit statistics were calculated. The goodness-of-fit statistics were monitored to ensure that the fit of the model never became inadequate. If all of the items with loading values of magnitude less than 0.05 on the postinstruction data were removed, then the process was redone using the preinstruction data. This back and forth was continued, incrementing by 0.05 in the magnitude of the loading values, until the goodness-of-fit statistics for the model became unacceptable.

This procedure resulted in the most parsimonious models possible while still having acceptable goodness-of-fit indices. An example of a model generated from this procedure can be found in Table 5.4 on the following page, whose details will be discussed in the results section (see Sec. 5.5.2 on page 147).

The goodness-of-fit statistics used in this study can be loosely placed into three categories: (i) parsimonious fit indexes, (ii) absolute fit indexes, and (iii) comparative fit indexes. The following is a brief discussion of some of these statistics, and a more detailed explanation can be found in Ref. [26].

Generally, as more parameters are added to a model the fit of the model will become better. Parsimonious fit indexes add a penalty to the goodness of fit in a manner related to the number of parameters used in the model to correct for a model having more parameters and another. Some indexes of this type are the root mean square error of approximation (RMSEA) and the M2 statistic [15, 54, 62]. The

Table 5.4: The 7-factor exploratory model generated through the iterative item removal process described in Sec. 5.4.3 on page 139. The related goodness-of-fit statistics for this model can be found in Table D.2 on page 202 for the exploratory and confirmatory samples of data.

F1	F2	F3	F4	F5	F6	F7
5	4	14	13	6	17	
11	15	21	29	8	25	
13	16	22	30	10	26	All Questions
18	28	23		17		
30		27		23		
				24		

RMSEA is bounded on bottom by 0 and is unbounded above. If the top of the 95% confidence interval of the RMSEA is below 0.06, then the model is said to have an acceptable fit [11]. The M2 statistic could be used to indicate acceptable model fit, but was instead used for parsimonious model selection. If two models have the same absolute and comparative goodness-of-fit values, then the model with the smaller M2 statistic is the preferred model as it is the most parsimonious between the two.

Absolute fit indexes are fit statistics that do not need to reference another model to be calculated. For example, the standardized root mean square of the residual (SRMSR) is an absolute fit index since it uses only the residuals of the model to determine goodness of fit. The SRMSR is bounded between 0 and 1, where a value of 0 indicates that all of the residuals were zero, and thus good fit. A model is said to have an acceptable fit if its SRMSR is below 0.08 [11, 41].

Lastly, comparative fit indexes are fit values calculated by comparing the fit of

the proposed model to a baseline model. A baseline model is one which makes no assumptions about the structure of the assessment and takes all of the parameters to be uncorrelated. The comparative fit index (CFI) and the Tucker-Lewis index (TLI) are two commonly used comparative fit indices. Both statistics are bounded on the bottom by 0 and on the top by 1. Values closer to 1 indicate good model fit, and values closer to 0 indicate that the proposed model and the baseline model fit the data with similar accuracies. Acceptable values for these statistics are above 0.90 or 0.95 depending on the desired model-fit strictness [11, 41].

5.4.4 Confirmatory multitrait item response theory

Confirmatory multitrait item response theory (CMIRT) is similar to the exploratory analysis described previously, the primary difference is that the models investigated in CMIRT are expert-proposed factors rather than data derived factors. Expert-proposed models are generated through the careful consideration of the conceptual content of the questions on the assessment and may be partially, or wholly, guided by exploratory results. These expert-proposed models are assumed to encapsulate the conceptual content within the assessment in a better manner than exploratory driven models because they are generated by experts of the field.

The models tested in this study were previously studied by us in Ref. [26] using CFA; see Tables 5.2 on page 133 and 5.3 on page 135 for descriptions of the models. In total 3 factor models for the FCI were investigated, 2 expert-proposed models and 1 model which resulted from exploratory factor analysis, and were found to adequately represent the factor structure of the FCI [26]. The exploratory model (SS5) was generated by Scott and Schumayer in 2012 in which they used exploratory factor analysis on 2150 postinstruction student responses from an algebra-based introductory physics I course [85]. They then reconfirmed the existence of this

structure by performing an EMIRT on the same set of data [82]. Another of the expert models analyzed was a slightly modified version of the factor model proposed by the creators of the FCI (HWS6) [38], and the last model was created using a combination of the creator's model and the EFA model from Scott and Schumayer (EW5) [26]. The factor structure and the qualitative meaning for each of these model's factors can be found in Tables 5.2 on page 133 and 5.3 on page 135, respectively. A thorough discussion and justification of these models is left to Refs. [26, 38, 85].

To perform CMIRT, one begins by specifying the expert model and then fits it against data from students who took the assessment being investigated. Once the model has been fit to the data, goodness-of-fit statistics of the model can be generated. All of this was done using a maximum likelihood estimator from the mirt package in R [14].

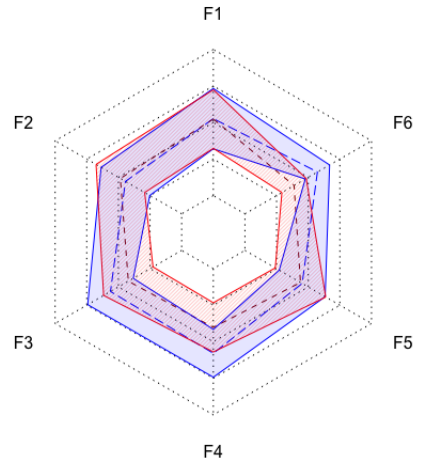
When analyzing the expert-proposed models, the covariances between factors were set to zero to make the student ability scores easier to interpret and to make the models as parsimonious as possible. Although, one would expect the concepts measured by the FCI to be related to one another, this study seeks factors that are independent from one another. To mode this, the factors were taken to be not correlated within the exploratory and confirmatory analysis. This interpretation of factor independence is used to analyze the results of the modified expert models in Sec. 5.6 on page 152.

5.4.5 Student ability radar plots

Once the modified expert-proposed models were confirmed to have a good fit with the preinstruction and postinstruction data, the linking procedure described in Sec. 5.4.2 on page 137 was used and the student abilities were estimated. This enabled the student ability scores to be estimated on the same ability scale for preinstruction

and postinstruction, which allowed for direct comparisons of the students' before and after instruction performances on the FCI.

Figure 5.3: Preinstruction (red) and postinstruction (blue) student ability estimations for the EW5 model. The short and long dashed lines represent the preinstruction and postinstruction medians, respectively. The shaded regions represent the the first and third quartiles of the students' postinstruction ability values for each of the traits. Similarly for the hashed regions and the preinstruction results. The outer- and innermost black dotted lines indicate $\theta = 2$ and $\theta = -2$, respectively. Notice only factor F3, F4, and F6 show growth from pretest to post-test. A description of the factors for the EW5 model can be found in Tables 5.2 on page 133 and 5.6 on page 149.



The resulting student abilities were plotted in radar plots using the "fmsb" package in R [70]. A radar plot is a graphical technique that allows for multivariate results to be displayed in a two-dimensional plot. For the plots presented in this study, the outermost dotted line corresponds to a θ of 2 and the innermost dotted line to -2. The student abilities of the sample for each trait are plotted along a spoke of the radar plot, with each spoke representing a single trait. An example of a preinstruction and postinstruction student abilities radar plot for the Eaton and

Willoughby 5-factor model can be found in Fig. 5.3 on the preceding page, where preinstruction results are displayed in red and postinstruction results are in blue. From the plots the traits on which the students experienced explanatory growth over the course of instruction become readily identifiable.

Table 5.5: Fit statistics for the expert-proposed factor models using the exploratory and confirmatory halves of the student response data. Since all three models fit both halves of the data equally well, the smallest M2 fit statistic (bold-faced) is used to identify the preferred model.

Exploratory half of the data							
Name	M2	RMSEA	RMSEA_5	RMSEA_95	SRMRS	TLI	CFI
EW5 Pre	**9548.885**	0.0446	0.04378	0.0453	0.04206	0.9543	0.9512
EW5 Post	**9225.783**	0.04376	0.0430	0.0445	0.0387	0.9596	0.9568
HWS6 Pre	10390.98	0.0464	0.0457	0.0472	0.0443	0.9501	0.9467
HWS6 Post	11346.57	0.0486	0.0478	0.0494	0.0391	0.9485	0.9449
SS5 Pre	111870.03	0.0499	0.0491	0.0507	0.0427	0.9474	0.9438
SS5 Post	12034.5891	0.0503	0.0495	0.0511	0.0427	0.9467	0.943
Confirmatory half of the data							
Name	M2	RMSEA	RMSEA_5	RMSEA_95	SRMRS	TLI	CFI
EW5 Pre	**10238.8**	0.0460	0.0453	0.0468	0.0428	0.9509	0.9475
EW5 Post	**8869.934**	0.0427	0.0419	0.0435	0.0395	0.9602	0.9574
HWS6 Pre	10390.98	0.0464	0.0457	0.0472	0.0443	0.9501	0.9467
HWS6 Post	11346.57	0.0486	0.0478	0.0494	0.0391	0.9485	0.9449
SS5 Pre	13035.23	0.0522	0.0515	0.0530	0.0498	0.9368	0.9325
SS5 Post	11263.53	0.0484	0.0477	0.0492	0.0438	0.9489	0.9453

5.5 Results

First the initial CMIRT results are discussed, followed by the subsequent EMIRT results. Lastly, the results for the CMIRT and the corresponding student ability radar plots are discussed.

5.5.1 Initial confirmatory multitrait item response theory Results

The unmodified expert-proposed factor models generated unacceptable model-fit statistics, with TLI and CFI values below 0.9, within the CMIRT frame work. This suggests something is missing from these models, which is causing them to inadequately model the students' response vectors. Thus, an EMIRT was performed to identify potential modifications that could be made to the expert models to increase their fit to the data without unintentionally adding unmotivated or unjustified features.

5.5.2 Exploratory multitrait item response theory

The factor structures for each of the resulting 3- through 7-factor exploratory models can be found in Table D.1 on page 201 and their goodness-of-fit statistics can be found in Table D.2 on page 202, both located in the Appendix. From Table D.2 on page 202, it can be seen that the models using 3 – 7 factors had acceptable model-fit values with the odd and even halves of the data. This suggests that the models that were generated from exploratory work are reliable and thus may be effectively modeling the population which the sample was drawn from. From the same table it can be seen that the models using 3 – 6 factor all have quite similar goodness-of-fit statistics. This is due to the manner in which the exploratory models were generated. Recall, the removal of items from traits was stopped only after the model-fit values became unacceptable. For these models, the SRMSR was the first fit value to become

unacceptable.

The utilization of 7 factors resulted in a model that did not have any issues with unacceptable fit statistics. The reason the iterative removal of items from factors was stopped for this model was to avoid factors with less than 3 items, which is not recommended for factor models [11]. From Table D.1 on page 201 it can be seen that the 7-factor model is more simplistic compared to the other exploratory models. Since the reduction of this model was not stopped due to fit concerns, the fit of this model passes even the strictest fit criteria for model goodness of fit.

The factor structure of the 7-factor exploratory model is actually a bi-factor model in that it contains a single, all-questions-included factor (sometimes called the general factor) with subfactor structure overlaid [30, 44, 100]. Coincidentally, the subfactor structure for the 7-factor model is almost identical to the proposed expert models from Ref. [26]. Comparing the composition of the factors for the 7-factor model and the expert models the following traits readily emerge:

(i) F1 - Identification of forces,

(ii) F2 - Newton's third law,

(iii) F3 - Newton's second law with kinematics,

(iv) F5 - Newton's first law with kinematics, and

(v) F6 - Superposition of forces or mixed methods.

The initial poor fit of the expert-proposed models and their similarities to the subfactors of the 7-factor model suggests that a general trait is necessary to properly model the FCI within a MIRT framework.

In short, a bi-factor model contains a general trait that is measured by all of the items on the assessment. Subtraits, or subfactors, are then specified on top of the general factor to account for residual variance shared by subsets of items on the assessment. A bi-factor model can appear when the assessment contains an

Table 5.6: The bi-factor expert-proposed models used in the CMIRT analysis. The related goodness-of-fit statistics for these models can be found in Table 5.5 on page 146 for the exploratory and confirmatory samples of data.

Hestenes, Wells, and Swackhamer 6-Factor Model (HWS6)							
F1	F2	F3	F4		F5	F6	F7
6	9	12	1	13	4	17	
7	22	14	2	18	15	25	
8	26	19	3	25	16	26	All Questions
10	27	20	5	30	28		
23		21	11				
24							

Eaton and Willoughby 5-Factor Model (EW5)							
F1		F2		F3	F4	F5	F6
6	20	9	21	5	4	17	
7	23	12	22	11	15	25	
8	24	14	27	13	16	26	All Questions
10		19		18	28		
				30			

Scott and Schumayer 5-Factor Model (SS5)							
F1		F2		F3	F4	F5	F6
6	12	19	22	5	4	16	
7	16	20	23	11	15	17	
8	24	21	27	13	28	25	All Questions
10	29			18			
				30			

overall scale (due to nonorthogonal factors) and orthogonal subscales [19, 80]. Thus, it makes sense that the exploratory analysis of the FCI suggests that expert models should incorporate a general trait since the factors on the FCI have been found to be

nonorthogonal in nature [85].

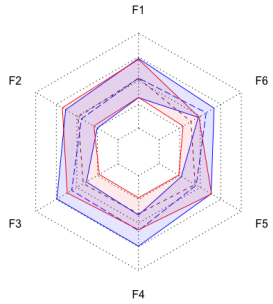
5.5.3 Confirmatory multitrait item response theory

Motivated by the results of the EMIRT analysis, the expert models were updated by adding a general trait, see Table 5.6 on the previous page. This adjustment resulted in acceptable goodness of fit for all three of the expert-proposed models with TLIs and CFIs all above 0.93 and RMSEAs and SRMSRs all below 0.06, see Table 5.5 on page 146.

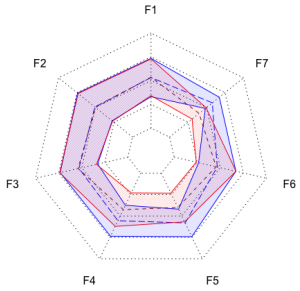
Table 5.7: Preinstruction (red) and postinstruction (blue) student ability estimations for the expert proposed models. The short and long dashed lines represent the preinstruction and postinstruction medians, respectively. The shaded regions represent the first and third quartiles of the students' postinstruction ability values for each of the traits. Similarly for the hashed regions and the preinstruction results. The outer- and innermost black dotted lines indicate $\theta = 2$ and $\theta = -2$, respectively. The factor compositions for these models can be found in Tables 5.2 on page 133 and 5.6 on page 149.

Exploratory half of data

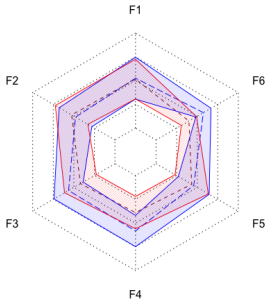
EW 5:



HWS 6:

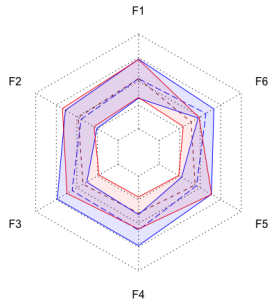


SS 5:

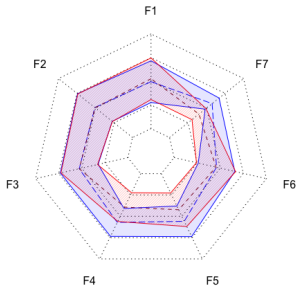


Confirmatory half of data

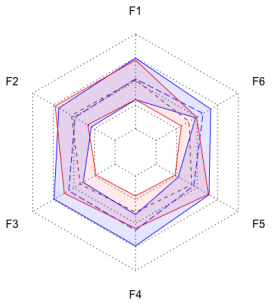
EW 5:



HWS 6:



SS 5:



5.6 Discussion

Research Question 1: Do the expert-proposed models tested in Ref. [26] adequately explain the data in a confirmatory multitrait item response theory framework?

The initial expert-proposed models used in Ref. [26] did not have acceptable goodness-of-fit statistics with the data in a CMIRT framework (TLIs and CFIs below 0.9).

Research Question 2: Using exploratory multitrait item response theory as a guide, can the expert-proposed models from Ref. [26] be modified to better fit the data, and what do these modifications suggest about the FCI?

The initially poor model fits and the results of the EMIRT analysis, suggest that the FCI may be best represented by a bi-factor model, when working in a MIRT statistical framework. For the FCI, the unidimensional scale represents the students' overall "Newtonian-ness" and the subscales represent the specific concepts probed by the instrument (Newton's third law, identification of forces, etc.). In conclusion, the expert models perform well if a general trait is included in the models.

Research Question 3: Provided an adequate confirmatory model is found, which of the models' factors demonstrate measurable student gain from preinstruction to postinstruction independent of the other factors?

Over the course of instruction, each of the expert model's preinstruction to postinstruction student ability radar plots show an improved student Newtonian-ness. This can be seen in Table 5.7 on the previous page, where there is outward motion on the general factor for each of the models (the last factor in each model). This implies that some of the students' improvement on the items is explained by a general increase in their overall Newtonian-ness. It should be noted that while attempting to model the FCI using just a unidimensional MIRT model acceptable goodness-of-fit statistics

could not be obtained. This suggests the need for a multidimensional model to help explain some of the remaining variance and thus improve the model-fit statistics.

Improved student ability scores along the unidimensional trait is expected since the students' aggregate total scores over the course of instruction improved on average. However, these results do not include student growth with respect to any of the specific concepts probed. Student growth along any of the other factors, which represent independent concepts, is thus information gained independent of what a unidimensional model would explain. For example, in Table 5.7 on page 151, the models all have the added general trait in the last factor of the radar plot. Thus, the growth along F3 and F4 in the EW5 model, F4 and F5 in the HWS6 model, and F3 and F4 in the SS5 model is information not captured by the unidimensional factors in each of the models. Traits with this kind of characteristic are defined as usable scales within the MIRT framework.

Regarding the subfactor structure, measurable gains were detected using the present data for students on the same two conceptual traits for all of the expert models, see Table 5.7 on page 151 in conjunction with Table 5.2 on page 133. These traits pertain to Newton's third law and identification of forces. This implies that student growth along these conceptual ideas could not be fully modeled using a unidimensional model. Although the other factors did explain some of the variance in the assessment, they did not help to explain what the students learned over the course of instruction independent from a general ability trait.

As a result of this analysis, the FCI can be thought of primarily as unidimensional assessment with subscales measuring concepts such as Newton's three laws, identification of forces, superposition of forces, and kinematics, depending on the expert model used. Of these subscales, this study identifies only 2 that are usable for identifying concepts students learned independently from the other traits on the

FCI: Newton's third law and identification of forces. Recall this idea of independence comes from the fact that the traits were all taken to have a correlation of zero with respect to one another. This implies that the traits can be taken to be independent of each other within the model. So within a MIRT perspective, if a researcher wanted to independently comment on students' gains in Newton's first and second laws and kinematics the FCI could not be used in its current form.

It should be noted that a model was investigated which included only the general trait, and the factors pertaining to Newton's third law and identification of forces, leaving out the other factors which did not exhibit student ability growth. When this model was fit against the student response data using CMIRT, unacceptable fit indices were generated. So, the other factors in the models are necessary to adequately fit student data. This suggests that the FCI is a 5 or 6 factor assessment, but only two of these factors were found in this study to exhibit student ability changes independent of the general, Newtonian-ness trait.

Since the results of this analysis come from modeling the FCI, they can be thought of as describing the characteristics of the FCI independent of the students. This is a critical feature of item response theory that item characteristics are independent of the student characteristics. This suggests that after proper modifications are made to the FCI some of the issues detected may be fixed. For example, changes could be made to make the questions measuring Newton's first and second laws to make them more independent from one another in an effort to separate these concepts from one another. This could result in better resolution of these factors, and thus allow for detectable student ability growth along these factors independent of the general trait.

5.7 Limitations

The results presented in this study were for a general population of introductory physics I students. Results based on more "pure" data may be slightly different and further research to check these results is suggested to strengthen the claims and utility of the results presented in this paper.

Since listwise deletion was used to clean the data used in this study, some bias may have been introduced into the results. Listwise deletion was used for this study for two reasons: (i) This study used the same original pool of data as Ref. [26], which used listwise deletion. To enable easier comparisons between the results of these studies listwise deletion was also used here. (ii) It has been found that regression results remain robust and relatively unbiased when using listwise deletion [2]. Since the 2PL model is similar to a logistic regression, the parameter estimations of MIRT should also be accurate. Further, standard errors found for regression models using listwise deleted data are found to still accurately estimate the true standard errors.

There could be other models not examined in this study that could produce a structure in which Newton's first and second laws and kinematics demonstrate student ability growth independent of the general unidimensional factor. However, considering the results of the EMIRT, these other models are unlikely to exist for the current form of the FCI since they would likely be significantly different from the discovered EMIRT models. This implies that these hypothetical models would not agree with the factors generated via student responses to the FCI and will thus not likely reflect an honest representation of the instrument.

5.8 Future Works

Future work will be to extend this analysis to other well-used conceptual assessments in PER, like the Force and Motion Conceptual Evaluation [97], the Conceptual Survey of Electricity and Magnetism [60], and the Brief Electricity and Magnetism Assessment [21]. This would help supply evidence for the existence of stable scales for these assessments, which would benefit concept specific interventions within these physical topics (i.e., laboratories or tutorials).

5.9 Conclusions

This study found that the original expert-proposed models for the FCI from Ref. [26] did not adequately fit the data within a MIRT framework. Suggestions for modifications to these models were found through the use of EMIRT,from which a bi-factor structure was discovered. Coincidentally, the 7-factor exploratory model had a striking resemblance to the proposed expert models analyzed in this study, with the notable inclusion of a general, unidimensional factor.

Since the FCI seeks to measure student Newtonian-ness by probing student understanding of concepts like Newton's third law or Force identification, a bi-factor model is reasonable. The general factor can be assumed to measure the Newtonian-ness of students, while the subfactors of the model measure the individual concepts used to probe the general trait. The validity of such a model was tested by adding a general factor to the expert models. The new expert bi-factor models were found to have acceptable model-fits with the data preinstruction and postinstruction. This suggests that the FCI is likely best modeled as an assessment which measures the general Newtonian-ness of students and then measures subabilities covering the concepts Newton's first law with kinematics, Newton's second law with kinematics,

Newton's third law, identification of forces, and superposition of forces.

From preinstruction and postinstruction student ability scores for these sub-factors, it was found that only two concepts consistently displayed student growth from independent of the general factor. These concepts were Newton's third law and identification of forces. This implies that within a MIRT framework the FCI may not be sensitive enough to measure the fine-grain growths of all the concepts contained on the assessment.

The lack of independent student ability growth found for some of these concepts is important to understand with respect to future curricula research. For instance, when examining the effectiveness of kinematics curricula or tutorials, using the FCI is not advised since it appears to be insensitive towards measuring independent student growth in kinematics. Researchers should consider choosing a different assessment if they are attempting to measure differences in students' abilities within specific concepts of force and motion, such as kinematics or Newton's second law. This study found that the 1995 version of the FCI could be used to measure students' overall Newtonian-ness, and independently measure students' understanding of Newton's third law and/or identification of forces.

Currently, some PER curricula research is preformed with conceptual or "in-house" instruments whose psychometric properties are not yet well understood. To ensure the validity and reproducibility of such studies, we encourage researchers to only use instruments whose psychometric properties are well understood. To enable single-concept specific research in PER, currently used instruments need to be better vetted statistically and qualitatively. The authors suggest that the future direction of PER involve the construction of single-conception assessments to avoid the inherent difficulties of creating a many-concept inventory which obeys a prespecified factor structure. Instruments of this nature would allow for more fine-grain curricula studies

to be performed with more confidence, precision, and accuracy than may be possible with the current assessments available within the field.

5.10 Acknowledgments

P.E. and S.W. would like to thank Physport for allowing us to use their data. Further P.E. and S.W. would like to thank Dr. Keith Johnson and David Russell for their insightful discussions and thoughtful comments when reading over the manuscript. This project was funded by the Montana State University Physics Department.

CHAPTER SIX

CONCLUSION

The current understanding of the instruments used in PER is severely lacking. However, the utilization of these instruments is unavoidable when performing education-oriented experiments. For a study to produce trustworthy results, extreme attention must be made when analyzing the collected data. This necessitates a detailed understanding of the psychometric functionality of the instruments used within these experiments, and by extension PER as a whole. These psychometric examinations should be performed prior to using any instruments outside of their validation studies. Incidentally, some instruments currently used in PER have undergone minimal analysis and most have not yet undergone strict validation studies. This kind of unintentional oversight could result in invalid statistics, and weak or incorrect conclusions for studies mistakenly using instruments that are actually not reliable and/or valid due to our currently poorly understanding of their functionality (see Chapter 3 as an example).

Currently, the preponderance of evidence suggests the overarching conclusions made by PER are likely valid. For example, the conclusion that active learning is generally superior to passive learning at all levels of education is well supported for most STEM fields. However, PER studies attempting to examine the effectiveness of single-concept interventions (e.g., tutorials, laboratories, and recitations) using multidimensional instruments may not have reported accurate statistical results (see Chapter 3 for details). The overall conclusions from these studies will likely remain the same, but the statistical significance of their conclusions will certainly change. Any changes in the statistics discussed in research could have significant impacts on

instructors who are attempting to identify the best teaching tools to use in their classrooms.

Within PER, there is currently a general lack of understanding of our multi-dimensional instruments' abilities to accurately resolve single conceptions. This makes testing single-concept interventions difficult, if not impossible. Additionally, the unfortunate lack of single-concept instruments further complicates this particular issue. The results of this dissertation, echoing the conclusions of Chapter 4 and Chapter 5, suggest the need for well-validated, single-concept instruments to allow for more fine-grain studies to be performed. These instruments do not currently exist for most topics in physics and would thus need to be constructed.

Until the time PER has the tools needed to perform fine-grained latent trait measurements, it is unavoidable that the multidimensional instruments currently in circulation will be used. If these instruments are not used, then curricula intervention research would have to be paused until single-concept tools are constructed. This could take on the order of 3-5 years to build a sufficient number of such instruments that have been rigorously psychometrically examined. Understandably, this is a largely unpopular option.

To be clear, the multidimensional instruments currently used in PER *can* assess the effectiveness of single-concept interventions *provided* the functionality of the instruments are well understood, and the corresponding student responses are analyzed properly. To perform these kinds of analyses, however, requires a good understanding of the psychometric functionality of these instruments, which does not currently exist for most tools being used in the field.

This dissertation, and the work contained within, helps to alleviate some of the aforementioned issues of the instruments being used in PER by addressing the following research question: How well is the foce concept inventory (FCI) functioning

at the distractor-level and at the factor-level? The conclusions gathered regarding this question only apply to the FCI, however, the tools and methods used can be immediately applied to any of the multiple-choice instruments currently being used in PER. This dissertation can be taken as part of a roadmap for how to properly assess the functionality of any multiple-choice instrument that is meant to be used for research.

Previous research regarding the psychometrics of the FCI make up a body of test- and item-level analyses and have found the FCI to be a well-functioning conceptual instrument. But, no previous examinations of the FCI have performed a distractor-level analysis. Thus, the general functionality of the distractors of the FCI was still unknown until the work within this dissertation was performed (Chapter 2).

Incidentally, most of these test- and item-level analyses were performed while assuming the FCI was a unidimensional (single-concept) instrument. This dissertation has shown that the FCI is in-fact multidimensional and should be modeled as such (see Chapter 3). Thus, the results of these previous studies should be reexamined.

Prior to this work, the multidimensional nature of the FCI had been examined within some research articles [82, 85, 86]. However, none of these articles formally demonstrated the need for a multidimensional model nor formally fit their multidimensional models for the FCI to student data in a confirmatory manner. The work presented in this dissertation builds upon this previous research by formally demonstrating the need for a multidimensional model to properly examine the FCI (Chapter 3), and uses new and previously proposed factor models to confirm the multidimensional structure of the FCI (Chapter 4 and Chapter 5).

The following is a brief breakdown and discussion of what each of the articles presented in this dissertation have added to the overall knowledge of the PER field. This present work has taken many steps in filling-in missing analyses and elucidating

the functionality of the FCI.

Chapter 2 on page 17 proposed a new partial credit score generating method that allowed for a distractor-level analysis of the FCI. From this analysis two questions were identified as likely not functioning as initially designed. Specifically, item 16 was identified as likely probing Newton's first and third laws when it was designed to only probe Newton's third law, and item 29 included the concept of buoyancy which the 1995 version of the FCI was designed to probe. The other items and options on the FCI were found to be functioning properly and likely do not need to be adjusted in future iterations of this instrument. These results can be used to help lead future efforts in updating this instrument and in the future generation of new and improved instruments.

The methodology proposed in Chapter 2 can be readily applied to any other multiple-choice instruments currently used in PER to lend an understanding of their distractor-level functionality. Additionally, this method could also be used to assess the scales of individual Likert-style questions to ensure student responses are being interpreted properly. Specifically, this methodology could be used to assess the assumption that the Likert-style response options represent an equally spaced discretization of the underlying continuous latent trait being measured. This could be assessed by generating the partial credit scores of the options and verifying their spacing from the scores.

Chapter 3, Chapter 4, and Chapter 5 supply a semi-complete multidimensional analysis of the FCI. The only missing portion of a full multidimensional investigation is the initial exploratory analysis. This exploratory analysis was included since it was already performed prior to this dissertation by Scott, Schumayer, and Gray [85]. These results were taken as a starting point for the multidimensional analysis performed in this dissertation.

Before the factor structure of the FCI was tested, the questions of its need for a multidimensional model was addressed. Chapter 3 presented a new methodology that tested for whether or not a unidimensional model is sufficient for examining a multiple-choice assessment. By considering the local dependence between items on the FCI, it was found that it should not be modeled using a unidimensional method. Instead, the FCI should be analyzed using multidimensional methods like factor analysis or multi-trait item response theory.

This methodology also showed that item chaining was likely not influencing students' responses to questions on the FCI. This implies that any grouping of items in correlation space is likely due to the presence of an underlying concept that links the items together and not due to superficial, superfluous correlations. These results implied that further analysis of the factor structure for the FCI were warranted.

The form the multidimensional model of the FCI should take could not be readily inferred from the result of Chapter 3. Fortunately, other studies can be used to inform a multidimensional model for the FCI. For instance, the exploratory factor analysis (EFA) model presented by Scott, Schumayer, and Gray made for an excellent starting point in developing a multidimensional model for the FCI [85]. Their model, in addition to two theoretical models, was tested using confirmatory factor analysis in Chapter 4. One of the theoretical models was generated from the test structure initially proposed by Hestenes *et. al.* (the creators of the FCI) in Ref. [38]. The other was developed by Eaton and Willoughby, in Ref. [26], as a hybrid of Scott, Schumayer, and Gray's EFA model and Hestenes *et. al.*'s model. It was found, unsurprisingly, that the EFA generated model best fit student data, but was not entirely expert-like in the way it grouped questions. Both Hestenes *et. al.*'s proposed model and the Eaton and Willoughby model were found to adequately fit student response data taken from the FCI. Between these two models, the Eaton and Willoughby model

was found to be the most stable and best explained the factors of the FCI.

The Eaton and Willoughby model proposed that the FCI was probing 5 Newtonian concepts. The first construct pertains to Newton's first law plus kinematics (FCI items 6, 7, 8, 10, 20, 23, 24). The second concept probes Newton's second law plus kinematics (FCI items 9, 12, 14, 19, 21, 22, 27). Specifically, kinematics within these concepts pertain to properly identifying the path an object would take, or how the speed of an object is changing for physical situations in which either Newton's first or second law apply. The third concept probed by the FCI is Newton's third law (FCI items 4, 14, 16, 28). However, it should be noted that item 16 has been identified as malfunctioning. Item 16 appears to be probing both Newton's third and first law, depending on how a student conceptually approaches the problem. This item will need to be updated/changed when the FCI is updated in the future. Force identification is the fourth concept investigated (FCI items 5, 11, 13, 18, 30), and Mixed conceptual methods is the fifth, and final, concept of the FCI (FCI item 17, 25, 26). Mixed conceptual methods pertains to problems in which students must use two or more of the other four concepts probed by the FCI to get the correct answer. Item 29 does not appear in this model since it was found to be malfunctioning in previous studies. The conceptual structure of this model closely resembles both the EFA model and the creators' model of the FCI.

Extending the confirmatory factor analysis performed in Chapter 4 to a multitrait item response theory framework, the fit of these factor models was investigated further in Chapter 5. It was found that all of the models tested previously required a minor modification. This modification took the form of the addition of a general factor, which is a factor that contains all of the questions on the assessment. This resulted in each of the examined models taking on a bifactor structure. Recall, a bifactor structure is essentially a unidimensional model with a multidimensional

substructure. In the case of the FCI, the general factor represents students' overall Newtonian-ness and the substructure represents student understanding of individual concepts independent of their general Newtonian understanding.

The presence of a general factor on the FCI suggests the previous analyses that assumed a unidimensional model for the FCI may not be as incorrect as initially anticipated. The results of these studies still will not be entirely accurate, but their errors may small enough that their conclusions could remain unchanged overall. This also helps to explain why a unidimensional student ability is almost perfectly correlated to the aggregate total score for students on the FCI [104].

Analyzing the average preinstruction to postinstruction multitrait student abilities determined for each of the examined factor models suggests the FCI can measure independent student gains along the general, Newtonian-ness factor, and gains along Newton's third law and the Force Identification factors. Thus, the FCI is capable of being used to assess curricular interventions targeting general Newtonian Mechanics and single-concept interventions targeting Newton's third law or Force Identification.

The results of this dissertation improve the current understanding of the functionality of the FCI. However, more analysis should be performed using additional techniques, like structural equation modeling, growth modeling, and other psychometric measurement theories. Structural equation modeling and growth modeling may be the next step in supplying an understanding of how the FCI functions from preinstruction to postinstruction.

These analyses should not be limited to the FCI, however, as many of the instruments currently used in PER are only minimally understood. The more analysis studies performed on the FCI and other PER instruments now, the more insight researchers will have into how to properly update these instruments in the

future. Additionally, the information gained from these studies will help inform future research projects that attempt to develop much needed single-concept instruments for PER.

REFERENCES CITED

[1] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman. New instrument for measuring student beliefs about physics and learning physics: The colorado learning attitudes about science survey. *Physical review special topics-physics education research*, 2(1):010101, 2006.

[2] P. D. Allison. *Missing data*, volume 136. Sage publications, 2001.

[3] J. Aslanides and C. Savage. Relativity concept inventory: Development, analysis, and results. *Physical Review Special Topics-Physics Education Research*, 9(1):010118, 2013.

[4] F. B. Baker and S.-H. Kim. *Item response theory: Parameter estimation techniques*. CRC Press, 2004.

[5] D. A. Bernstein. Does active learning work? a good question, but not the right one. *Scholarship of Teaching and Learning in Psychology*, 4(4):290, 2018.

[6] Y. M. Bishop, S. E. Fienberg, and W. Paul. *Holland. Discrete multivariate analysis: theory and practice*. The MIT Press, 1975.

[7] R. D. Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51, 1972.

[8] R. D. Bock and I. Moustaki. 15 item response theory in a general framework. 26:469 – 513, 2006.

[9] A. H. Bowker. A test for symmetry in contingency tables. *Journal of the american statistical association*, 43(244):572–574, 1948.

[10] E. Brewe, J. Bruun, and I. G. Bearden. Using module analysis for multiple choice responses: A new method applied to force concept inventory data. *Physical Review Physics Education Research*, 12:020131, Sep 2016.

[11] T. A. Brown. *Confirmatory Factor Analysis for Applied Research*. 2 edition, 2015.

[12] M. W. Browne and R. Cudeck. Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2):230–258, 1992.

[13] R. C. MacCallum, M. W. Browne, and H. M. Sugawara. Power analysis and determination of sample size for covariance structure modeling. 1:130–149, 06 1996.

[14] R. P. Chalmers et al. mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48(6):1–29, 2012.

[15] F. Chen, Y. Liu, T. Xin, and Y. Cui. Applying the m2 statistic to evaluate the fit of diagnostic classification models in the presence of attribute hierarchies. *Frontiers in psychology*, 9, 2018.

[16] W.-H. Chen and D. Thissen. Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3):265–289, 1997.

[17] V. P. Coletta and J. A. Phillips. Interpreting fci scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics*, 73(12):1172–1182, 2005.

[18] R. J. de Ayala. *The Theory and Practice of Item Response Theory*.

[19] C. E. DeMars. A tutorial on interpreting bifactor model scores. *International Journal of Testing*, 13(4):354–378, 2013.

[20] R. F. DeVellis. Classical test theory. *Medical care*, pages S50–S59, 2006.

[21] L. Ding, R. Chabay, B. Sherwood, and R. Beichner. Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical review special Topics-Physics education research*, 2(1):010105, 2006.

[22] P. Eaton, K. Johnson, B. Frank, and S. Willoughby. Classical test theory and item response theory comparison of the brief electricity and magnetism assessment and the conceptual survey of electricity and magnetism. *Physical Review Physics Education Research*, 15(1):010102, 2019.

[23] P. Eaton, K. Johnson, and S. Willoughby. Generating a growth-oriented partial credit grading model for the force concept inventory. *Physical Review Physics Education Research*, 15:020151, Dec 2019.

[24] P. Eaton, K. Vavruska, and S. Willoughby. Exploring the preinstruction and postinstruction non-newtonian world views as measured by the force concept inventory. *Physical Review Physics Education Research*, 15(1):010123, 2019.

[25] P. Eaton and S. Willoughby. Identifying a preinstruction to postinstruction factor model for the force concept inventory within a multitrait item response theory framework. *Physical Review Physics Education Research*, 16:010106, Jan 2020.

[26] P. Eaton and S. D. Willoughby. Confirmatory factor analysis applied to the force concept inventory. *Physical Review Physics Education Research*, 14:010124, Apr 2018.

[27] A. Elby. Helping physics students learn how to learn. *American Journal of Physics*, 69(S1):S54–S64, 2001.

[28] L. R. Fabrigar and D. T. Wegener. *Exploratory Factor Analysis: Understanding Statistics*. Oxford University Press, inc., 2012.

[29] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23):8410–8415, 2014.

[30] R. D. Gibbons and D. R. Hedeker. Full-information item bi-factor analysis. *Psychometrika*, 57(3):423–436, 1992.

[31] R. R. Hake. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1):64–74, 1998.

[32] R. Hambleton and R. Jones. Comparison of classical test theory and item response theory and their applications to test development, instructional topics in educational measurement series 16, 2012.

[33] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig. Dividing the force concept inventory into two equivalent half-length tests. *Physical Review Special Topics-Physics Education Research*, 11(1):010112, 2015.

[34] P. Heller and D. Huffman. Interpreting the force concept inventory: A reply to hestenes and halloun. *The Physics Teacher*, 33(8):503–503, 1995.

[35] C. Henderson. Common concerns about the force concept inventory. *The Physics Teacher*, 40(9):542–547, 2002.

[36] D. Hestenes and I. Halloun. Interpreting the force concept inventory: A response to march 1995 critique by huffman and heller. *The Physics Teacher*, 33(8):502–502, 1995.

[37] D. Hestenes and M. Wells. A mechanics baseline test. *The physics teacher*, 30(3):159–166, 1992.

[38] D. Hestenes, M. Wells, and G. Swackhamer. Force concept inventory. *The Physics Teacher*, 30(3):141–158, 1992.

[39] C. R. Houts and M. C. Edwards. The performance of local dependence measures with psychological data. *Applied Psychological Measurement*, 37(7):541–562, 2013.

[40] C. R. Houts and M. C. Edwards. Comparing surface and underlying local dependence levels via polychoric correlations. *Applied psychological measurement*, 39(4):293–302, 2015.

[41] L. Hu and P. M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55, 1999.

[42] D. Huffman and P. Heller. What does the force concept inventory actually measure? *The Physics Teacher*, 33(3):138–143, 1995.

[43] M. Ishimoto, G. Davenport, and M. C. Wittmann. Use of item response curves of the force and motion conceptual evaluation to compare japanese and american students? views on force and motion. *Physical Review Physics Education Research*, 13(2):020135, 2017.

[44] K. G. Jöreskog. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2):183–202, 1969.

[45] T. Kline. *Classical Test Theory: Assumptions, Equations, Limitations, and Item Analyses*. SAGE Publications, Inc., Thousand Oaks, California, 2005.

[46] M. J. Kolen and R. L. Brennan. *Test equating, scaling, and linking: Methods and practices*. Springer Science & Business Media, 2014.

[47] K. D. Kubinger. On artificial results due to using factor analysis for dichotomous variables. *Psychology Science*, 25(1):106–110, May 2003.

[48] D. R. L. Implications for measurement and evaluation from the trends of science education. *Science Education*, 60(2):199–209, 1980.

[49] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef. The puzzling reliability of the force concept inventory. *American Journal of Physics*, 79(9):909–912, 2011.

[50] D. N. Lawley and A. E. Maxwell. *Factor analysis as a statistical method [by] D.N. Lawley and A.E. Maxwell*. Butterworths London, 1963.

[51] W.-C. Lee and J.-C. Ban. A comparison of irt linking procedures. *Applied Measurement in Education*, 23(1):23–48, 2009.

[52] Y. H. Li and R. W. Lissitz. An evaluation of the accuracy of multidimensional irt linking. *Applied Psychological Measurement*, 24(2):115–138, 2000.

[53] W. Lin, Q. Jiahe, and L. YiHsuan. Exploring alternative test form linking designs with modified equating sample size and anchor test length. *ETS Research Report Series*, 2013(1):i–17.

[54] Y. Liu, W. Tian, and T. Xin. An application of m 2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, 41(1):3–26, 2016.

[55] F. M. Lord. The relation of test score to the trait underlying the test. *Educational and psychological measurement*, 13(4):517–549, 1953.

[56] F. M. Lord and M. R. Novick. *Statistical theories of mental test scores*. IAP, 2008.

[57] K. J. Louis, B. J. Ricci, and T. I. Smith. Determining a hierarchy of correctness through student transitions on the fmce. In *Physics Education Research Conference 2018*, PER Conference, Washington, DC, August 1-2 2018.

[58] C. Magno. Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, 1(1):1–11, 2009.

[59] D. P. Maloney. Rule-governed approaches to physics-newton's third law. *Physics Education*, 19(1):37, 1984.

[60] D. P. Maloney, T. L. O?Kuma, C. J. Hieggelke, and A. Van Heuvelen. Surveying students? conceptual knowledge of electricity and magnetism. *American Journal of Physics*, 69(S1):S12–S23, 2001.

[61] J. D. Marx and K. Cummings. Normalized change. *American Journal of Physics*, 75(1):87–91, 2007.

[62] A. Maydeu-Olivares and H. Joe. Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71(4):713, 2006.

[63] E. Mazur and R. C. Hilborn. Peer instruction: A user's manual. *Physics Today*, 50:68, 1997.

[64] S. B. McKagan. Physport, 2011.

[65] R. L. McKinley and M. D. Reckase. An extension of the two-parameter logistic model to the multidimensional latent space. 1983.

[66] S. Monroe and L. Cai. Estimation of a ramsay-curve item response theory model by the metropolis–hastings robbins–monro algorithm. *Educational and Psychological Measurement*, 74(2):343–369, 2014.

[67] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley. Testing the test: Item response curves and test quality. *American Journal of Physics*, 74(5):449–453, 2006.

[68] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker. An item response curves analysis of the force concept inventory. *American Journal of Physics*, 80(9):825–831, 2012.

[69] G. A. Morris, P. Walter, S. Skees, and S. Schwartz. Transition matrices: A tool to assess student learning and improve instruction. *The Physics Teacher*, 55(3):166–169, 2017.

[70] M. Nakazawa. *fmsb: Functions for Medical Statistics Book with some Demographic Data*, 2018. R package version 0.6.3.

[71] M. L. Nering and R. Ostini. *Handbook of polytomous item response theory models*. Taylor & Francis, 2011.

[72] J. Nissen, R. Donatello, and B. Van Dusen. Missing data and bias in physics education research: A case for using multiple imputation. *Physical Review Physics Education Research*, 15(2):020106, 2019.

[73] P. Osteen. An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research*, 1(2):66–82, 2010.

[74] R. J. Patz and B. W. Junker. A straightforward approach to markov chain monte carlo methods for item response models. *Journal of educational and behavioral Statistics*, 24(2):146–178, 1999.

[75] M. Planinic, L. Ivanjek, and A. Susac. Rasch model based analysis of the force concept inventory. *Physical Review Physics Education Research*, 6:010103, Mar 2010.

[76] M. Prince. Does active learning work? a review of the research. *Journal of engineering education*, 93(3):223–231, 2004.

[77] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[78] S. Ramlo. Validity and reliability of the force and motion conceptual evaluation. *American Journal of Physics*, 76(9):882–886, 2008.

[79] P. Ramsden. *Learning to teach in higher education*. Routledge, 2003.

[80] S. P. Reise, T. M. Moore, and M. G. Haviland. Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment*, 92(6):544–559, 2010.

[81] Y. Rosseel. "lavaan: An r package for structural equation modeling". *Journal of Statistical Software*, 48(2):1 – 36, 2012.

[82] T. F. Scott and D. Schumayer. Students' proficiency scores within multitrait item response theory. *Physical Review Physics Education Research*, 11:020134, Nov 2015.

[83] T. F. Scott and D. Schumayer. Conceptual coherence of non-newtonian worldviews in force concept inventory data. *Physical Review Physics Education Research*, 13:010126, May 2017.

[84] T. F. Scott and D. Schumayer. Central distractors in force concept inventory data. *Physical Review Physics Education Research*, 14(1):010106, 2018.

[85] T. F. Scott, D. Schumayer, and A. R. Gray. Exploratory factor analysis of a force concept inventory data set. *Physical Review Physics Education Research*, 8:020105, Jul 2012.

[86] M. R. Semak, R. D. Dietz, R. H. Pearson, and C. W. Willis. Examining evolving performance on the force concept inventory using factor analysis. *Physical Review Physics Education Research*, 13:010103, Jan 2017.

[87] C. Singh and D. Rosengrant. Multiple-choice test of energy and momentum concepts. *American Journal of Physics*, 71(6):607–617, 2003.

[88] T. Smith, K. Gray, K. Louis, B. Ricci, and N. Wright. Showing the dynamics of student thinking as measured by the fmce. In *Proceedings, Physics Education Research Conference 2017*, page 380, 2018.

[89] T. I. Smith, K. J. Louis, I. Ricci, J. Bartholomew, and N. Bendjilali. Quantitatively ranking incorrect responses to multiple-choice questions using item response theory. *arXiv preprint arXiv:1906.00521*, 2019.

[90] T. I. Smith and M. C. Wittmann. Applying a resources framework to analysis of the force and motion conceptual evaluation. *Physical Review Special Topics-Physics Education Research*, 4(2):020101, 2008.

[91] R. P. Springuel, M. C. Wittmann, and J. R. Thompson. Applying clustering to statistical analysis of student reasoning about two-dimensional kinematics. *Physical Review Special Topics-Physics Education Research*, 3(2):020107, 2007.

[92] J. Stewart, M. Miller, C. Audo, and G. Stewart. Using cluster analysis to identify patterns in students? responses to contextually different conceptual problems. *Physical Review Special Topics-Physics Education Research*, 8(2):020112, 2012.

[93] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart. Multidimensional item response theory and the force concept inventory. *Physical Review Physics Education Research*, 14(1):010137, 2018.

[94] Y. Suh and D. M. Bolt. Nested logit models for multiple-choice item response data. *Psychometrika*, 75(3):454–473, 2010.

[95] D. Thissen, L. Steinberg, and A. R. Fitzpatrick. Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, 26(2):161–176, 1989.

[96] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx. Comparing the force and motion conceptual evaluation and the force concept inventory. *Physical review special topics-Physics education research*, 5(1):010105, 2009.

[97] R. K. Thornton and D. R. Sokoloff. Assessing student learning of newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *american Journal of Physics*, 66(4):338–352, 1998.

[98] R. Traub. A priori considerations in choosing an item response model. *Applications of item response theory*, 57:70, 1983.

[99] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell. Gender fairness within the force concept inventory. *Physical Review Physics Education Research*, 14(1):010103, 2018.

[100] L. R. Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23(2):111–136, 1958.

[101] A. L. Van den Wollenberg. Two new test statistics for the rasch model. *Psychometrika*, 47(2):123–140, 1982.

[102] J. Von Korff, B. Archibeque, K. A. Gomez, T. Heckendorf, S. B. McKagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell. Secondary analysis of teaching methods in introductory physics: A 50 k-student study. *American Journal of physics*, 84(12):969–974, 2016.

[103] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye. *Probability & statistics for engineers & scientists*. Prentice Hall, 2012.

[104] J. Wang and L. Bao. Analyzing force concept inventory with item response theory. *American Journal of Physics*, 78(10):1064–1070, 2010.

[105] P. Wattanakasiwich, P. Taleab, M. D. Sharma, and I. D. Johnston. Construction and implementation of a conceptual survey in thermodynamics. *International Journal of Innovation in Science and Mathematics Education (formerly CAL-laborate International)*, 21(1), 2013.

[106] J. Wells, R. Henderson, J. Stewart, G. Stewart, J. Yang, and A. Traxler. Exploring the structure of misconceptions in the force concept inventory with modified module analysis. *Physical Review Physics Education Research*, 15:020122, Sep 2019.

[107] M. Wilson. Detecting and interpreting local item dependence using a fannily of rasch models. *Applied Psychological Measurement*, 12(4):353–364, 1988.

[108] M. C. Wittmann and K. E. Black. Visualizing changes in pretest and post-test student responses with consistency plots. *Physical Review Special Topics-Physics Education Research*, 10(1):010114, 2014.

[109] Y. Xiao, K. Koenig, J. Han, Q. Liu, J. Xiong, and L. Bao. Test equity in developing short version conceptual inventories: A case study on the conceptual survey of electricity and magnetism. *Physical Review Physics Education Research*, 15(1):010122, 2019.

[110] J. Yang, C. Zabriskie, and J. Stewart. Multidimensional item response theory and the force and motion conceptual evaluation. *Physical Review Physics Education Research*, 15:020141, Nov 2019.

[111] L. Yao and K. Boughton. Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement*, 46(2):177–197, 2009.

[112] J.-i. Yasuda and M.-a. Taniguchi. Validating two questions in the force concept inventory with subquestions. *Physical Review Special Topics-Physics Education Research*, 9(1):010113, 2013.

[113] J.-i. Yasuda, H. Uematsu, and H. Nitta. Validating a japanese version of the force concept inventory. *Lat. Am. J. Phys. Educ. Vol*, 1(89):6, 2012.

[114] W. M. Yen. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2):125–145, 1984.

[115] W. M. Yen. Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement*, 30(3):187–213, 1993.

APPENDICES

APPENDIX A

COPY OF THE FORCE CONCEPT INVENTORY

1. Two metal balls are the same size but one weights twice as much as the other. The balls are dropped from the roof of a single story building at the same instant of time. The time it takes the balls to reach the group will be:

   (A) about half as long for the heavier ball as for the lighter one.

   (B) about half as long for the lighter ball as for the heavier one.

   (C) about the same for both balls.

   (D) considerably less for the heavier ball, but not necessarily half as long.

   (E) considerably less for the lighter ball, but not necessarily half as long.

2. Two metal balls of the previous problem roll off a horizontal table with the same speed. In this situation:

   (A) both balls hit the floor at approximately the same horizontal distance from the bas of the table.

   (B) the heavier ball hits the floor at about half the horizontal distance from the base of the table as does the lighter one.

   (C) the lighter ball hits the floor at about half the horizontal distance from the base of the table as does the heavier one.

   (D) the heavier ball hits the floor considerably closer to the base of the table than the lighter ball, but not necessarily at half the horizontal distance.

   (E) the lighter ball hits the floor considerably closer to the base of the table than the lighter ball, but not necessarily at half the horizontal distance.

3. A stone dropped from the roof of a single story building to the surface of the earth:

   (A) reaches a maximum speed quite a soon after release and then falls at a constant speed thereafter.

   (B) speeds up as it falls because the gravitational attraction gets considerably stronger as the stone gets closer to the earth.

   (C) speeds up because of an almost constant force of gravity acting upon it.

   (D) falls because of the natural tendency of all objects to rest on the surface of the earth.

   (E) falls because of the combined effects of the force of gravity pushing it downward and the force of the air pushing it downwards.

4. A large truck collides head-on with a small compact car. During the collision:

   (A) the truck exerts a greater amount of force on the car than the car exerts on the truck.

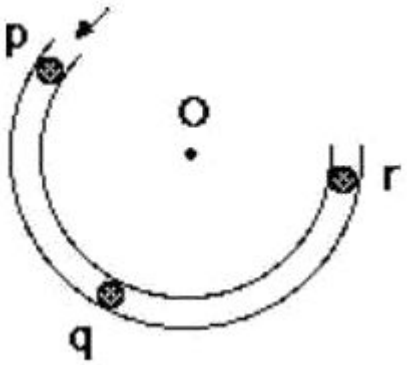
   (B) the car exerts a greater amount of force on the truck than the truck exerts on the car.

   (C) neither exerts a force on the other, the car gets smashed simply because it gets in the way of the truck.

   (D) the truck exerts a force on the cr, but the car does not 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.

The accompanying figure shows a frictionless channel in the shape of a segment of a circle with center at "O". the channel has been anchored to a frictionless horizontal table top. You are looking down at the table. Forces exerted by the air are negligible. A ball is shot at high speed into the channel at "p" and exits at "r."



5. Consider the following distinct forces:

   1. A downward force of gravity
   2. A force exerted by the channel pointing from q to O.
   3. A force in the direction of motion.
   4. A force pointing from O to q.

Which of the above forces is (are) acting on the ball when it is within the frictionless channel at position "q"?

(A) 1 only.

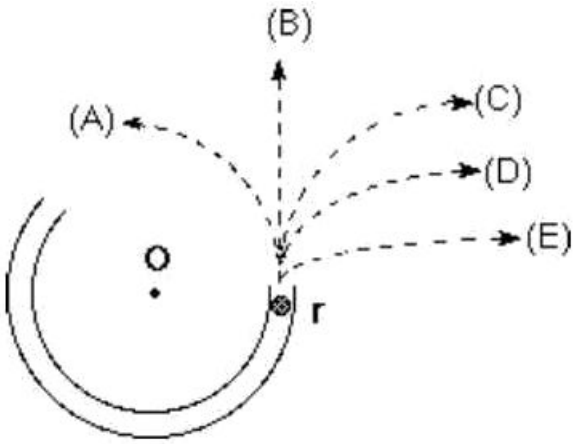
(B) 1 and 2.

(C) 1 and 3.

(D) 1, 2, and 3.

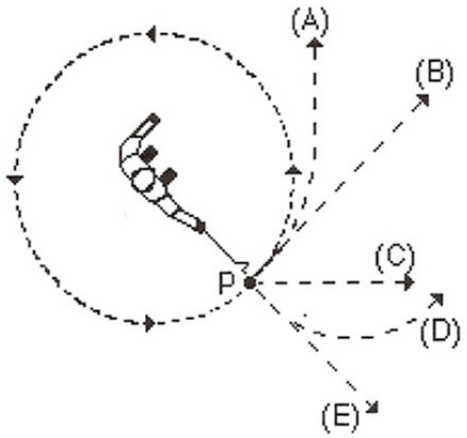
(E) 1, 3, and 4.

6. Which path in the figure [below] would the ball most likely follow after it exits the channel at "r" and moves across the frictionless table?



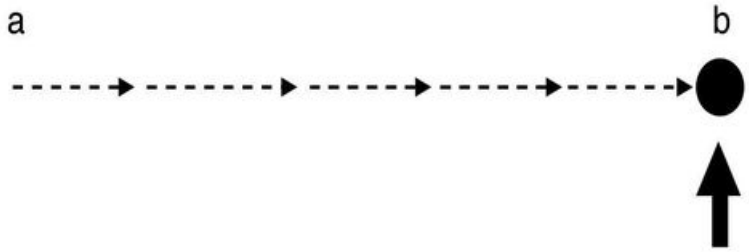
7. A steel ball is attached to a string and is swung in a circular path in a horizontal plane as illustrated in the accompanying figure. At the point P indicated in the figure, the string suddenly breaks near the ball.

   If these events are observed from directly above as in the figure, which path would the ball most closely follow after the string breaks?

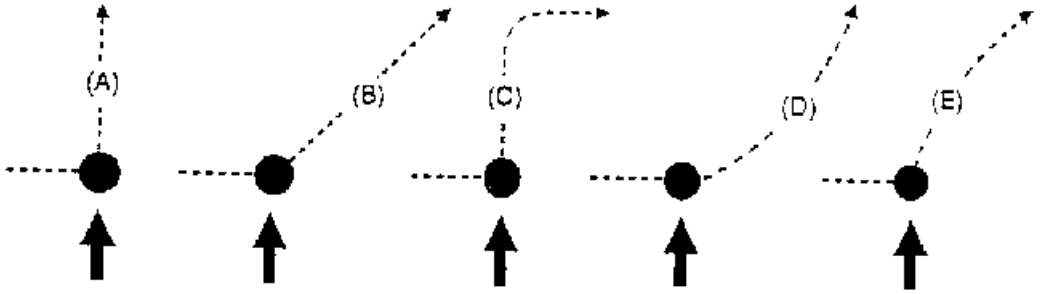


Use the statement and figure below to answer the next four questions (8 through 11).

The figure depicts a hockey puck sliding with constant speed v_0 in a straight line from point "a" to point "b" on a frictionless horizontal surface. Forces exerted by the air are negligible. You are looking down on the puck. When the puck reaches point "b," it receives a swift horizontal kick in the direction of the heavy printed arrow. Had the puck been at rest at point "b," then the kick would have set the puck in horizontal motion with speed v_k in the direction of the kick.



8. Which of the paths would the puck most closely follow after receiving the kick?



9. The speed of the puck just after it receives the kick is:

   (A) equal to the speed "v_0" it had before is received the kick.

   (B) equal to the speed "v_k" resulting from the kick and independent of the speed "v_0".

   (C) equal to the arithmetic sum of the speeds "v_k" and "v_k".

   (D) smaller than either of the speeds"v_k" or "v_k".

   (E) greater than either of the speeds"v_k" or "v_k", but less than the arithmetic sum of these two speeds.

10. Along the frictionless path you have chosen in question 8, the speed of the uck after receiving the kick:

    (A) is constant.

    (B) continuously increases.

    (C) continuously decreases.

    (D) increases for a while and decreased thereafter.

    (E) is constant for a while and decreases thereafter.

11. Along the frictionless path you have chosen in question 8, the main force(s) acting on the puck after receiving the kick is (are):

    (A) a downward force of gravity.

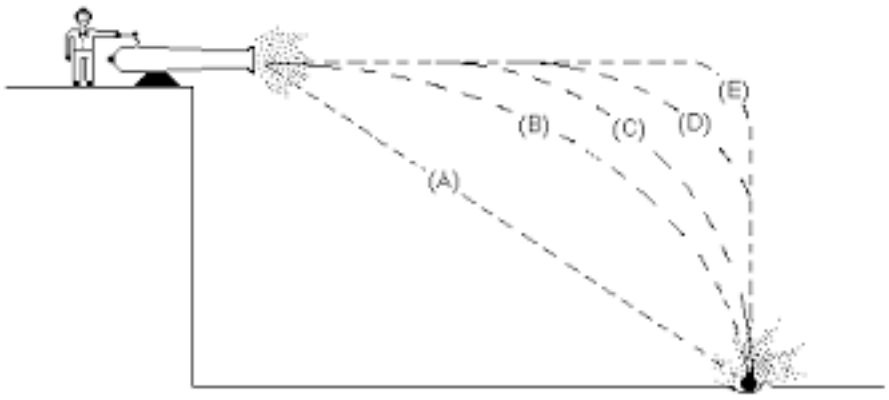
    (B) a downward force of gravity, and a horizontal force in the direction of motion.

    (C) a downward force of gravity, an upward force exerted by the surface, and a horizontal force in the direction of motion.

    (D) a downward force of gravity, and an upward force exerted by the surface.

    (E) none. (No forces act on the puck.)

12. A ball is fired by a cannon from the top of a cliff as shown in the figure below. Which s the paths would the cannon ball most closely follow?



13. A boy throws a steel ball straight up. Consider the motion of the ball only after it has left the boy's hand but before it touches the ground, and assume that forces exerted by the air are negligible. For these conditions, the force(s) acting on the ball is (are):

    (A) a downward force of gravity along with a steadily decreasing upward force.

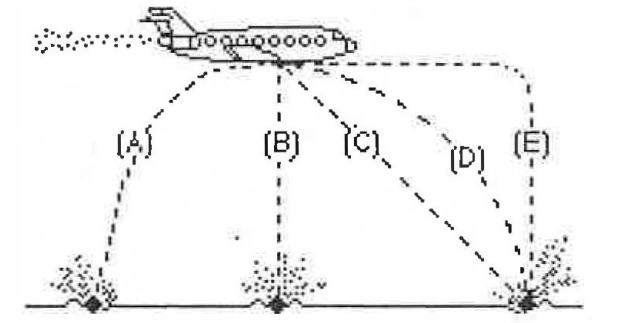
    (B) a steadily decreasing upward force from the moment it leaves the boy's hand until it reaches its highest point; on the way down there is a steadily increasing downward force of gravity as the object gets closer to earth.

    (C) an almost constant downward force of gravity along with an upwards force that steadily decreases until the ball reaches its highest point; on the way down there is only a constant downward force of gravity.

    (D) an almost constant downward force of gravity only.

    (E) none of the above. The ball falls back to ground because of its natural tendency to rest on the surface of the earth.

14. A bowling ball accidentally falls out of the cargo of an airliner as it flies along in a horizontal direction.
    As observed by a person standing on the ground and viewing the plane as in the figure at the right, which path would the bowling ball most closely follow after leaving the airplane?



Use the statement and the figure below to answer the next two questions (15 through 16).

A large truck breaks down out on the road and receives a push back into town by a small compact car as shown in the figure below.

15. While the car, still pushing the truck, is speeding up to get to cruising speed:

    (A) the amount of force with which the car pushes on the truck is equal to that with which the truck pushes back on the car.

    (B) the amount of force with which the car pushes on the truck is smaller than that with which the truck pushes back on the car.

    (C) the amount of force with which the car pushes on the truck is greater than that with which the truck pushes back on the car.

    (D) the car's engine is running so the car pushes against the truck, but the truck's engine is not running so the truck cannot push back against the car. The truck is pushed forward simply because it is in the way of the car.

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

16. After the car reaches the constant cruising speed at which its driver wishes to push the truck:

    (A) the amount of force with which the car pushes on the truck is equal to that with which the truck pushes back on the car.

    (B) the amount of force with which the car pushes on the truck is smaller than that with which the truck pushes back on the car.

    (C) the amount of force with which the car pushes on the truck is greater than that with which the truck pushes back on the car.

    (D) the car's engine is running so the car pushes against the truck, but the truck's engine is not running so the truck cannot push back against the car. The truck is pushed forward simply because it is in the way of the car.

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

17. An elevator is being lifted up an elevator shaft at a constant speed by a steel cable as shown in the figure below. All frictional effects are negligible. In this situation, forces on the elevator are such that:

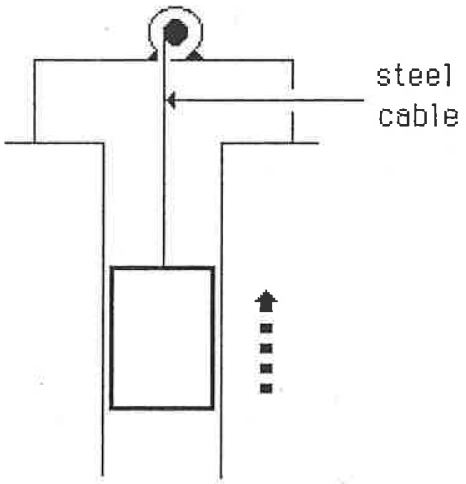
    (A) the upward force by the cable is greater than the downward force of gravity.

    (B) the upward force by the cable is equal to the downward force of gravity.

    (C) the upward force by the cable is smaller than the downward force of gravity.

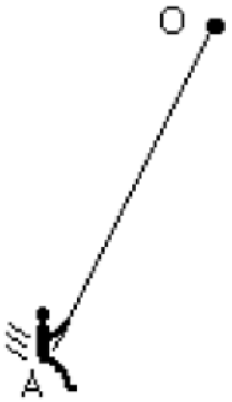
    (D) the upward force by the cable is greater than the sum of the downward force of gravity and a downward force due to the air.

    (E) none of the above. (The elevator goes up because the cable is being shortened, not because an upward force is exerted on the elevator by the cable).



Elevator going up at constant speed

18. The figure below shows a boy swinging on a rope, starting at a point higher than A.



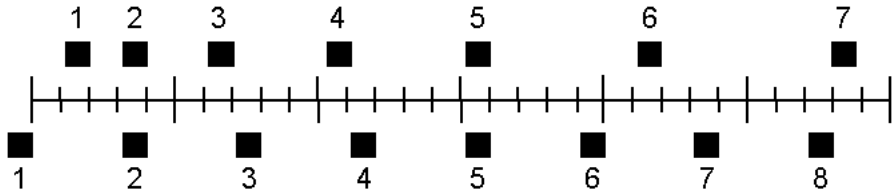
Consider the following distinct forces:

1. A downward force of gravity.
2. A force exerted by the rope pointing from A to O.
3. A force in the direction of the boy's motion.
4. A force pointing from O to A.

Which of the above forces is (are) acting on the boy when he is at position A?

(A) 1 only.
(B) 1 and 2.
(C) 1 and 3.
(D) 1, 2, and 3.
(E) 1, 3, and 4.

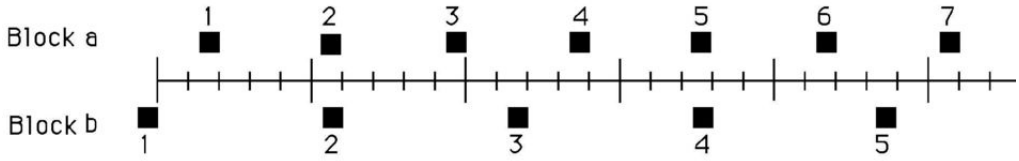
19. The position of two blocks at successive 0.20-second time intervals are represented by the numbered squares in the figure below. The blocks are moving toward the right.



Do the blocks ever have the same speed?

(A) No.
(B) Yes, at instant 2.
(C) Yes, at instant 5.
(D) Yes, at instants 2 and 5.
(E) Yes, at some time during the interval 3 to 4.

20. The position of two blocks at successive 0.20-second time intervals are represented by the numbered squares in the figure below. The blocks are moving toward the right.



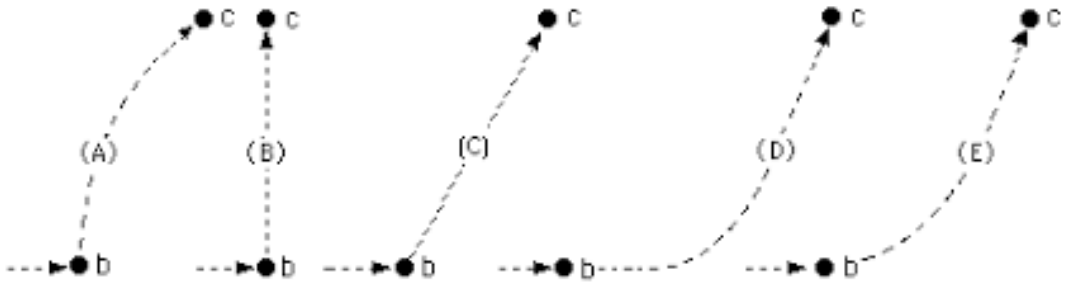
The acceleration of the blocks are related as follows:

(A) The acceleration of "a" is greater than the acceleration of "b".
(B) The acceleration of "a" equals the acceleration of "b". Both accelerations are greater than zero.
(C) The acceleration of "b" is greater than the acceleration of "a".
(D) The acceleration of "a" equals the acceleration of "b". Both accelerations are zero.
(E) Not enough information is given to answer the question.

Use the statement and figure below to answer the next four questions (21 through 24).

A rocket drifts sideways in outer space from point "a" to point "b" as shown below. The rocket is subjected to no outside forces. Starting at position "b", the rocket's engine is turned on and produces a constant thrust (force on the rocket) at right angles to the line "ab". The constant thrust is maintained until the rocket reaches a point "c" in space.

21. Which path below best represents the path of the rocket between points "b" and "c"?



22. As the rocket moves from position "b" to position "c", its speed is:

    (A) constant.

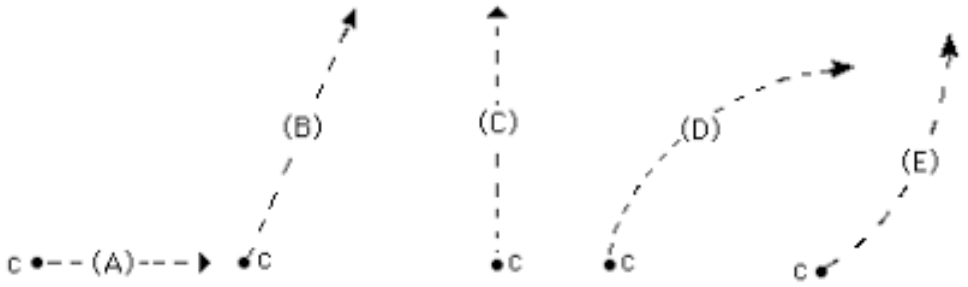
    (B) continuously increasing.

    (C) continuously decreasing.

    (D) increasing for a while and constant thereafter.

    (E) constant for a while and decreasing thereafter.

23. At point "c" the rocket's engine is turned off and the thrust immediately drops to zero. Which of the paths below will the rocket follow beyond point "c"?



24. Beyond position "c" the speed of the rocket is:

(A) constant.

(B) continuously increasing.

(C) continuously decreasing.

(D) increasing for a while and constant thereafter.

(E) constant for a while and decreasing thereafter.

25. A woman exerts a constant horizontal force on a large box. As a result, the box moves across a horizontal floor at a constant seed "v_0"." The constant horizontal force applied by the woman:

    (A) has the same magnitude as the weight of the box.

    (B) is greater then the weight of the box.

    (C) has the same magnitude as the total force which resists the motion of the box.

    (D) is greater then the total force which resists the motion of the box.

    (E) is greater that either the weight of the box or the total force which resists its motion.

26. If the woman in the previous question doubles the constant horizontal force that the exerts on the box to push it on the same horizontal floor, the box them moves:

    (A) with a constant speed that is double the speed "v_0" in the previous question.

    (B) with a constant speed that is grater then the speed "v_0" in the previous question, by not necessarily twice as great.

    (C) for a while with a speed that is constant and greater than the speed "v_0" in the previous question, then with a speed that increases thereafter.

    (D) for a while with an increasing speed, then with a constant speed thereafter.

    (E) with a continuously increasing speed.

27. If the woman in question 25 suddenly stops applying a horizontal force to the box, then the box will:

    (A) immediately come to a stop.

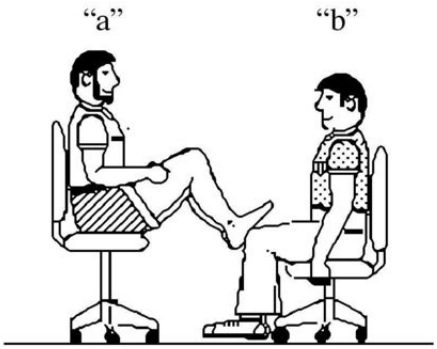
    (B) continue moving at a constant speed for a while and then slow to a stop.

    (C) immediately start slowing to a stop.

    (D) continue as a constant speed.

    (E) increase its speed for a while and then start slowing to a stop.

28. In the figure [below], student "a" has a mass of 95 kg and student "b" has a mass of 77 kg. They sit in identical office chairs facing each other.



Student "a" places hie bare feet on the knees of student "b", as shown. Student "a" then suddenly pushes outward with his feet, causing both chairs to move.

During the push and while the students are still touching each other:

(A) neither student exerts a force on the other.
(B) student "a" exerts a force on student "b", but "b" does not exert and force on "a".
(C) each student exerts a force on the other, but "b" exerts the larger force.
(D) each student exerts a force on the other, but "a" exerts the larger force.
(E) each student exerts the same amount of force on the other.

29. An empty office chair is at rest on a floor. Consider the following force:

1. A downward force on gravity.
2. An upward force exerted by the floor.
3. A net downward force exerted by the air.

Which of the forces is (are) acting on the office chair?

(A) 1 only.
(B) 1 and 2.
(C) 2 and 3.
(D) 1, 2, and 3.
(E) none of the forces. (Since the chair is at rest there are no forces acting upon it.)

30. Despite a very strong wind, a tennis player manages to hit a tennis ball with her racquet so that the ball passes over the net and lands in her opponent's court.
Consider the following forces:

    1. A downward force of gravity.
    2. A force by the "hit".
    3. A force exerted by the air.

    Which of the above forces is (are) acting on the tennis ball after it has heft contact with the racquet and before it touches the ground?

    (A) 1 only.
    (B) 1 and 2.
    (C) 1 and 3.
    (D) 2 and 3.
    (E) 1, 2, and 3.

APPENDIX B

TABLES AND FIGURES FOR CHAPTER ONE

Table B.1: Two-parameter logistics nominal response model's item parameters given as mean (standard deviation) for the 1000 uniformly sampled classes of 10 000 students. The ak_1 and b_1 values are taken to be zero in the model to set the measurement scale of the other parameters. The last three columns give the relative distance of the correct option from the nearest incorrect option in ak space, as given by the inverted key item estimations. The standard deviations for $\Delta ak_{\text{inv. Cor}}$ were calculated using $\sigma = \sqrt{\sigma^2_{\text{Correct Option}} + \sigma^2_{\text{Closest Distractor}}}$, where σ is the standard deviation.

Q	a	d	ak_1	ak_2	ak_3	ak_4	b_1	b_2	b_3	b_4	$ak_{\text{inv. Cor}}$	$ak_{\text{inv. 1st Dis}}$	$\Delta ak_{\text{inv. Cor}}$
1	1.137(0.027)	1.905(0.026)	0(0)	-0.319(0.061)	0.037(0.056)	-0.316(0.103)	0(0)	-0.863(0.066)	-0.188(0.055)	-1.937(0.112)	-1.098(0.038)	-0.047(0.057)	1.051(0.069)
2	0.937(0.019)	0.028(0.017)	0(0)	-0.33(0.06)	-0.131(0.028)	0.045(0.061)	0(0)	-1.678(0.048)	-0.116(0.024)	-1.961(0.048)	0(0)	0.842(0.054)	0.842(0.054)
3	1.161(0.023)	0.801(0.019)	0(0)	-0.699(0.042)	-0.569(0.064)	-0.706(0.073)	0(0)	-1.087(0.042)	-2.222(0.064)	-2.208(0.071)	-0.921(0.025)	0(0)	0.921(0.025)
4	0.963(0.023)	1.037(0.02)	0(0)	-0.923(0.131)	-1.788(0.194)	-0.795(0.172)	0(0)	-3.935(0.152)	-5.258(0.279)	-4.121(0.182)	-0.911(0.023)	0(0)	0.911(0.023)
5	1.426(0.027)	-0.343(0.02)	0(0)	-1.116(0.066)	-0.387(0.056)	-0.757(0.062)	0(0)	0.419(0.046)	1.547(0.035)	0.619(0.041)	-0.956(0.045)	0(0)	0.956(0.045)
6	1.516(0.039)	2.518(0.037)	0(0)	-0.401(0.093)	-1.083(0.151)	-0.694(0.238)	0(0)	-1.812(0.105)	-3.654(0.211)	-3.804(0.297)	-1.421(0.041)	0(0)	1.421(0.041)
7	1.16(0.028)	1.727(0.026)	0(0)	0.332(0.058)	-0.617(0.098)	0.545(0.057)	0(0)	-0.069(0.058)	-1.957(0.121)	0.134(0.056)	-1.405(0.047)	-0.542(0.057)	0.863(0.074)
8	1.299(0.027)	0.988(0.021)	0(0)	-1.788(0.17)	0.153(0.05)	0.16(0.042)	0(0)	-4.47(0.251)	-0.364(0.045)	-0.058(0.04)	-1.381(0.035)	-0.165(0.05)	1.216(0.061)
9	0.98(0.021)	0.179(0.018)	0(0)	-0.197(0.054)	0.347(0.051)	-0.134(0.07)	0(0)	1.309(0.046)	1.678(0.042)	-0.181(0.061)	-1.026(0.045)	-0.299(0.045)	0.727(0.064)
10	1.969(0.039)	1.904(0.032)	0(0)	-0.144(0.093)	-0.845(0.093)	-0.612(0.094)	0(0)	0.634(0.082)	0.338(0.087)	-0.092(0.085)	0(0)	1.53(0.077)	1.53(0.077)
11	1.711(0.032)	0.327(0.023)	0(0)	-2.331(0.095)	-1.544(0.065)	-0.394(0.082)	0(0)	-1.648(0.079)	0.756(0.034)	-0.995(0.052)	-0.687(0.046)	0(0)	0.687(0.046)
12	1.261(0.032)	2.028(0.03)	0(0)	-0.129(0.113)	-1.115(0.155)	-1.249(0.172)	0(0)	2.361(0.1)	-0.782(0.171)	-1.755(0.206)	-1.031(0.12)	0(0)	1.031(0.12)
13	2.561(0.05)	0.498(0.029)	0(0)	-0.4(0.087)	0.57(0.081)	-1.481(0.328)	0(0)	0.343(0.086)	2.236(0.073)	-4.097(0.45)	-2.933(0.09)	-0.52(0.078)	2.413(0.119)
14	1.116(0.021)	0.285(0.018)	0(0)	0.373(0.034)	0.672(0.04)	-1.923(0.246)	0(0)	-0.271(0.03)	-0.442(0.032)	-6.367(0.402)	-1.406(0.029)	-0.656(0.037)	0.75(0.047)
15	0.48(0.017)	-0.024(0.014)	0(0)	0.908(0.058)	-1.138(0.143)	-1.457(0.348)	0(0)	2.748(0.055)	-2.372(0.194)	-4.562(0.542)	0(0)	0.389(0.018)	0.389(0.018)
16	1.335(0.031)	1.963(0.028)	0(0)	0.584(0.102)	-0.322(0.147)	2.237(0.145)	0(0)	2.045(0.123)	-0.871(0.191)	1.554(0.131)	0(0)	0.143(0.047)	0.143(0.047)
17	1.212(0.023)	-0.64(0.019)	0(0)	-1.875(0.126)	-0.504(0.044)	-0.566(0.078)	0(0)	-4.955(0.165)	-2.108(0.037)	-3.301(0.066)	-1.134(0.023)	0(0)	1.134(0.023)
18	1.674(0.03)	0.155(0.021)	0(0)	-0.563(0.12)	0.038(0.105)	-0.304(0.108)	0(0)	0.327(0.093)	2.518(0.073)	1.614(0.078)	-1.577(0.1)	0(0)	1.577(0.1)
19	1.27(0.025)	0.774(0.019)	0(0)	-0.386(0.08)	0.151(0.06)	-0.063(0.04)	0(0)	-1.257(0.078)	-0.506(0.053)	0.963(0.038)	-1.243(0.04)	-0.153(0.06)	1.09(0.072)
20	1.17(0.024)	0.334(0.018)	0(0)	0.647(0.053)	0.288(0.035)	0.125(0.076)	0(0)	-0.343(0.043)	0.901(0.034)	-1.52(0.067)	-0.759(0.048)	0(0)	0.759(0.048)
21	0.991(0.021)	-0.253(0.017)	0(0)	0.38(0.055)	1.057(0.052)	0.72(0.057)	0(0)	0.89(0.055)	2.073(0.049)	1.026(0.054)	-1.394(0.036)	-0.625(0.033)	0.769(0.049)
22	1.313(0.024)	0.439(0.02)	0(0)	-1.116(0.093)	-0.373(0.039)	-1.482(0.116)	0(0)	-2.867(0.096)	-0.669(0.032)	-3.841(0.14)	-1.155(0.025)	0(0)	1.155(0.025)
23	1.38(0.026)	0.466(0.02)	0(0)	0.421(0.044)	0.007(0.046)	-0.005(0.073)	0(0)	0.958(0.04)	0.503(0.043)	-0.952(0.066)	-1.58(0.043)	-0.419(0.044)	1.161(0.062)
24	1.703(0.037)	2.034(0.033)	0(0)	0.144(0.093)	-0.431(0.132)	-0.403(0.107)	0(0)	1.734(0.093)	-0.673(0.147)	-0.185(0.114)	0(0)	1.605(0.039)	1.605(0.039)
25	1.399(0.027)	-0.584(0.021)	0(0)	-0.442(0.081)	0.911(0.07)	0.078(0.077)	0(0)	0.177(0.079)	2.769(0.06)	0.826(0.066)	-1.977(0.068)	-0.758(0.063)	1.219(0.093)
26	1.898(0.036)	-1.393(0.027)	0(0)	0.225(0.027)	-0.844(0.075)	0.379(0.037)	0(0)	-0.024(0.021)	-2.763(0.078)	-0.763(0.027)	-2.054(0.038)	-0.397(0.036)	1.657(0.052)
27	1.043(0.023)	0.833(0.019)	0(0)	-0.274(0.038)	0.216(0.072)	-0.696(0.165)	0(0)	-0.925(0.037)	-1.997(0.054)	-3.726(0.171)	-0.992(0.025)	-0.235(0.069)	0.757(0.073)
28	1.437(0.029)	1.084(0.023)	0(0)	-0.38(0.123)	0.285(0.111)	0.687(0.103)	0(0)	-0.005(0.138)	1.116(0.12)	3.144(0.111)	-2.039(0.112)	-0.712(0.111)	1.327(0.158)
29	0.274(0.02)	1.222(0.017)	0(0)	-0.28(0.116)	1.277(0.071)	0.344(0.11)	0(0)	-0.674(0.147)	2.683(0.078)	-0.451(0.119)	-1.553(0.086)	-1.526(0.088)	0.027(0.123)
30	1.495(0.028)	-0.219(0.02)	0(0)	-0.936(0.067)	-0.297(0.077)	-0.624(0.055)	0(0)	0.01(0.052)	-0.394(0.054)	1.807(0.035)	-0.951(0.051)	0(0)	0.951(0.051)

Figure B.1: Average item scores for questions 1–15 pre and post-test. Averages from dichotomous scoring are in red and the partial credit scoring item averages are in blue.

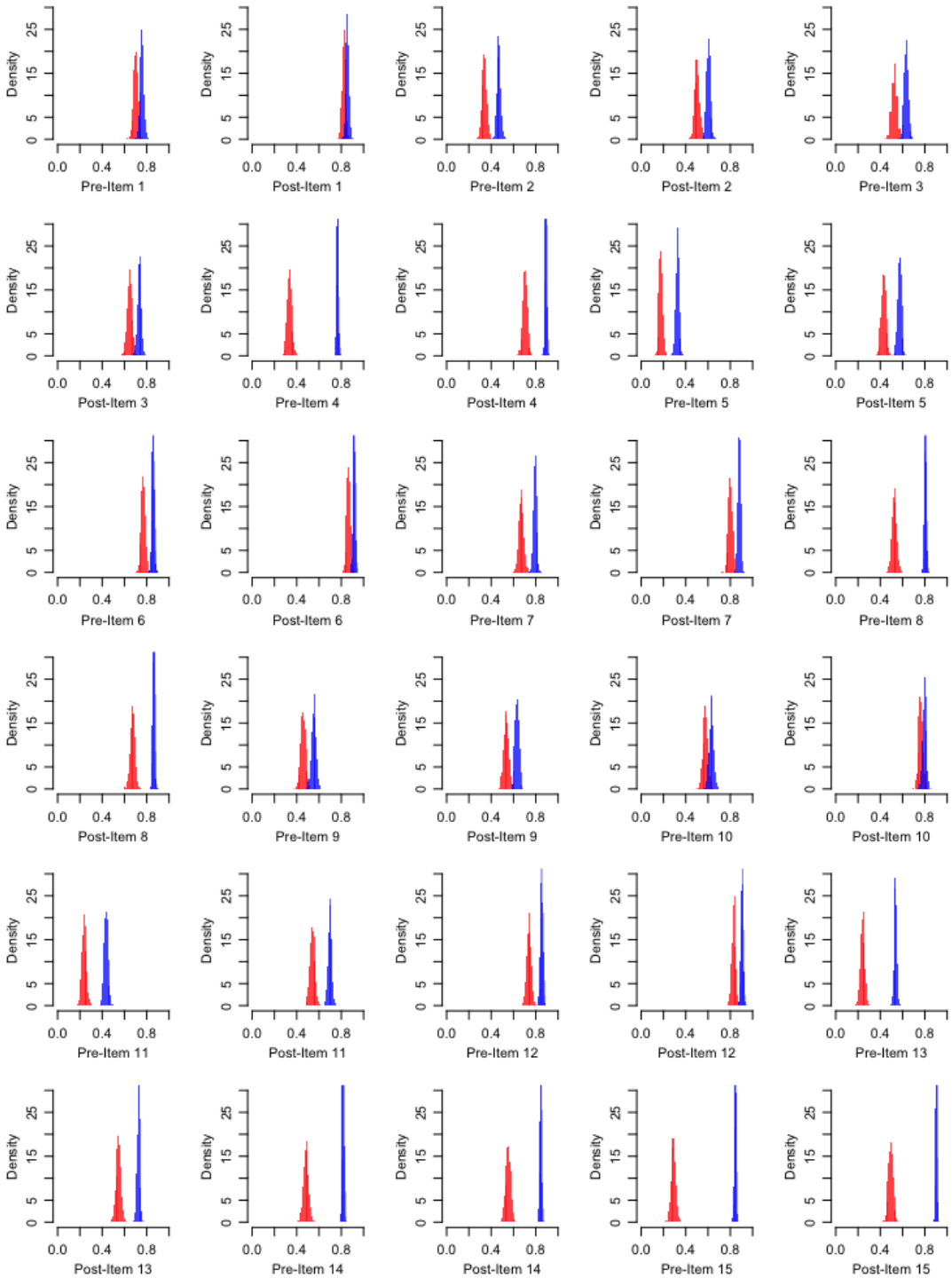
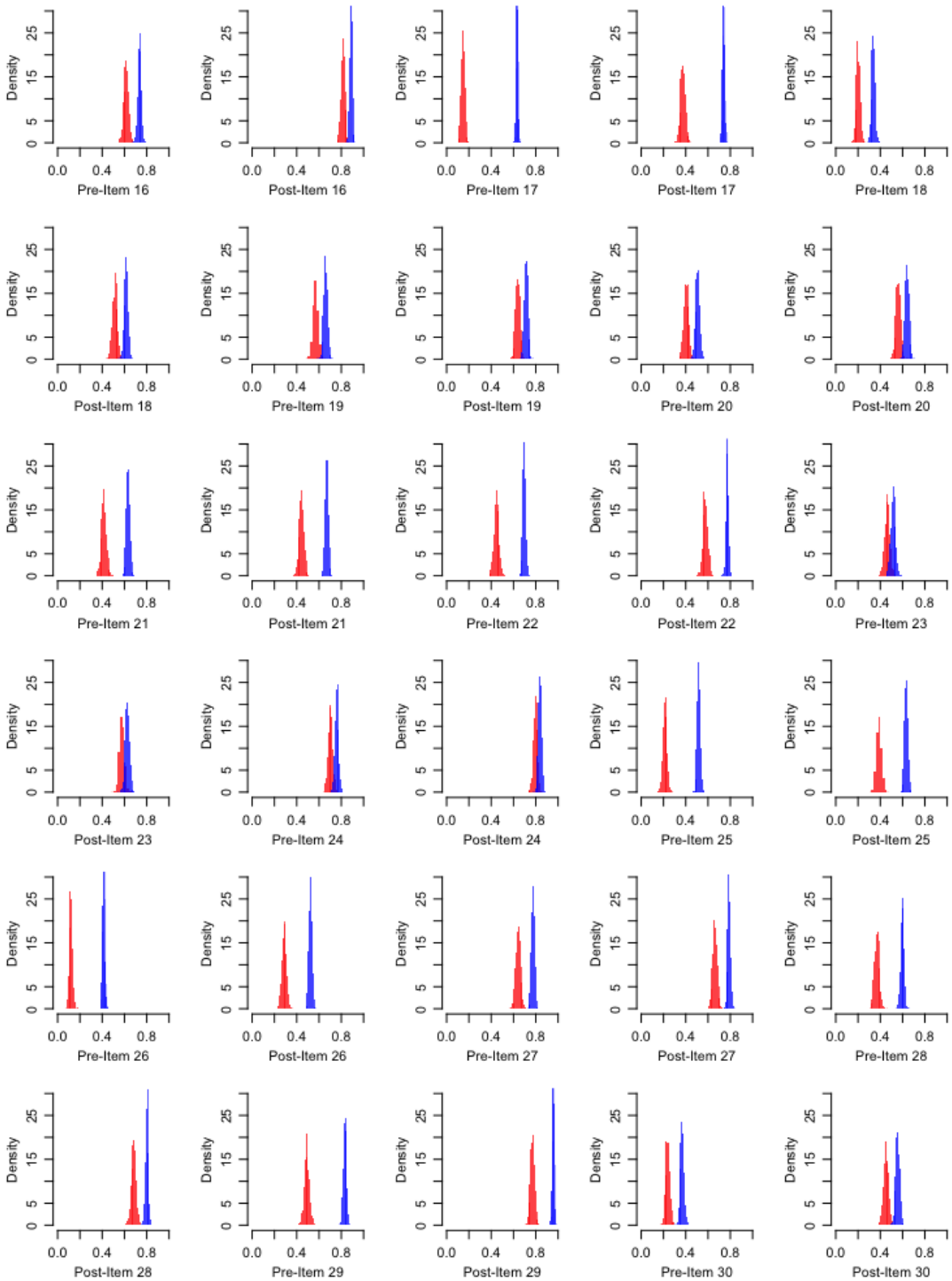


Figure B.2: Average item scores for questions 16–30 pre and post-test. Averages from dichotomous scoring are in red and the partial credit scoring item averages are in blue.



APPENDIX C

TABLES AND FIGURES FOR CHAPTER TWO

Table C.1: The proposed cutoff values for ULD weights ranging from 0 - 5 in steps of 0.1. The boldfaced rows are the cutoff values used in this study.

ULD Weight	V_{χ^2}	V_{G^2}	r_{tet}	ULD Weight	V_{χ^2}	V_{G^2}	r_{tet}
0.0	**0.06422541**	**0.06416194**	**0.6208568**	**2.5**	**0.6504**	**0.6926841**	**0.9266251**
0.1	0.06136577	0.06153779	0.6119237	2.6	0.6670881	0.7114637	0.9345929
0.2	0.0709306	0.07140207	0.6121322	2.7	0.6784003	0.7237656	0.9361162
0.3	0.07672111	0.07761015	0.6265379	2.8	0.6889258	0.7355418	0.9427226
0.4	0.1024029	0.1043197	0.64817	2.9	0.6903578	0.7390411	0.9400761
0.5	0.121908	0.1245156	0.6568833	3	0.7075131	0.758766	0.950447
0.6	0.1618948	0.1662142	0.7019055	3.1	0.718474	0.7698705	0.9497939
0.7	0.1963584	0.2024259	0.718134	3.2	0.7289987	0.782845	0.9540501
0.8	0.2225444	0.2308232	0.7275791	3.3	0.7325532	0.7882925	0.9600743
0.9	0.2697526	0.2807468	0.7590138	3.4	0.7427828	0.7985638	0.9568945
1.0	0.2967564	0.308299	0.7645016	3.5	0.7569794	0.8135756	0.9639189
1.1	0.3179208	0.3323178	0.7705089	3.6	0.7588925	0.8177181	0.9604786
1.2	0.3660675	0.3824436	0.803789	3.7	0.7606699	0.8218524	0.9633361
1.3	0.3946269	0.4136427	0.8178026	3.8	0.7775624	0.841649	0.970911
1.4	0.4319518	0.4510915	0.8366347	3.9	0.7776252	0.8406047	0.9661712
1.5	**0.4547008**	0.**4783952**	**0.8512132**	4	0.8012364	0.8640076	0.97319
1.6	0.4720992	0.4958578	0.8509541	4.1	0.793261	0.85692	0.9706856
1.7	0.5038353	0.5333389	0.8728725	4.2	0.8036369	0.8689481	0.9729531
1.8	0.5178476	0.5460599	0.8735163	4.3	0.8055976	0.8742733	0.9780525
1.9	0.5425084	0.5737948	0.8860569	4.4	0.812301	0.8788273	0.9741356
2.0	**0.5638722**	**0.5958849**	**0.8945882**	4.5	0.8119624	0.8809874	0.9758418
2.1	0.5758467	0.6113034	0.8954174	4.6	0.8120548	0.8810052	0.9747251
2.2	0.5935914	0.6303532	0.9071281	4.7	0.8253878	0.8953174	0.9777068
2.3	0.6238864	0.6630311	0.9173902	4.8	0.8263691	0.9017414	0.9817862
2.4	0.6323589	0.6715236	0.9184202	4.9	0.8359095	0.9096283	0.9798809
				5.0	0.835386	0.909959	0.9816036

APPENDIX D

TABLES AND FIGURES FOR CHAPTER FOUR

Table D.1: The 3- through 7-factor EMIRT models. Goodness-of-fit statistics of these models for the exploratory and confirmatory halves of the data can be found in Table D.2 on the next page.

7-Factor Exploratory Model						
F1	F2	F3	F4	F5	F6	F7
5	4	14	13	6	17	
11	15	21	29	8	25	
13	16	22	30	10	26	All Questions
18	28	23		17		
30		27		23		
				24		

6-Factor Exploratory Model								
F1	F2	F3	F4	F5		F6		
2	2	1	3	1	17	1	11	21
5	4	2	9	2	19	3	12	22
11	15	8	14	3	20	4	13	23
13	16	10	19	8	22	5	14	24
17	28	23	21	9	25	6	15	26
18	29	24	22	10	26	7	16	27
22			23	13	28	8	18	28
25			26	14	30	9	19	30
26			27	16		10	20	
30								

5-Factor Exploratory Model								
F1		F2	F3	F4		F5		
2	18	4	9	2	22	1	11	21
3	20	15	19	4	25	2	12	22
5	22	16	20	13	26	3	13	23
10	25	20		15	28	5	14	24
11	26	28		16	30	6	16	26
13	29	29		17		7	17	27
17	30					8	18	28
						9	19	30
						10	20	

4-Factor Exploratory Model							
F1		F2	F3		F4		
2	18	4	2	22	1	11	20
3	20	15	4	25	2	12	21
5	22	16	13	26	3	13	22
8	25	18	15	28	4	14	23
10	26	21	16	30	6	15	24
11	28	22	17		7	16	26
13	29	28			8	17	27
16	30	29			9	18	28
17					10	19	30

3-Factor Exploratory Model								
F1			F2			F3		
1	12	22	1	16		1	12	23
2	13	23	2	18		2	13	24
3	14	25	4	20		3	14	26
5	16	26	6	28		6	16	27
6	17	28	7	29		7	18	28
8	18	30	15			8	19	29
9	19					9	20	30
10	20					10	21	
11	21					11	22	

Table D.2: Fit statistics for the exploratory factor models using the exploratory and confirmatory halves of the student response data.

Exploratory half of data							
Name	M2	RMSEA	RMSEA_5	RMSEA_95	SRMRS	TLI	CFI
7-factor exploratory model pre	7602.8436	0.0395	0.0387	0.0403	0.0338	0.9641	0.9616
7-factor exploratory model post	6821.7249	0.0373	0.0365	0.0381	0.0345	0.9707	0.9687
6-factor exploratory model pre	11473.7269	0.0491	0.0483	0.0498	0.0748	0.9447	0.9408
6-factor exploratory model post	11760.0204	0.0497	0.0489	0.0505	0.0719	0.9479	0.9443
5-factor exploratory model pre	12182.2793	0.0506	0.0498	0.0514	0.0794	0.9411	0.937
5-factor exploratory model post	12539.887	0.0514	0.0506	0.0521	0.076	0.9443	0.9405
4-factor exploratory model pre	10701.1697	0.0473	0.0465	0.0481	0.0653	0.9485	0.945
4-factor exploratory model post	11536.8648	0.0492	0.0484	0.05	0.0707	0.949	0.9454
3-factor exploratory model pre	12728.7833	0.0518	0.051	0.0525	0.0753	0.9383	0.9341
3-factor exploratory model post	16485.3091	0.0592	0.0584	0.0599	0.0739	0.9262	0.9211
Confirmatory half of data							
Name	M2	RMSEA	RMSEA_5	RMSEA_95	SRMRS	TLI	CFI
7-factor exploratory model pre	8149.188	0.0408	0.0401	0.0416	0.0345	0.9614	0.9587
7-factor exploratory model post	6603.938	0.0365	0.0357	0.0373	0.0355	0.9709	0.9689
6-factor exploratory model pre	11812.43	0.0496	0.0489	0.0504	0.0739	0.9430	0.9391
6-factor exploratory model post	11374.96	0.0487	0.0479	0.0494	0.0698	0.9483	0.9448
5-factor exploratory model pre	12404.81	0.0509	0.0501	0.0517	0.0771	0.9400	0.9359
5-factor exploratory model post	11978.75	0.0500	0.0492	0.0508	0.0740	0.9455	0.9417
4-factor exploratory model pre	10921.81	0.0477	0.0469	0.0484	0.0634	0.9475	0.9438
4-factor exploratory model post	11335.75	0.0486	0.0478	0.0494	0.0686	0.9485	0.9450
3-factor exploratory model pre	12874.1	0.0519	0.0511	0.0527	0.0734	0.9377	0.9334
3-factor exploratory model post	15830.35	0.0578	0.0570	0.0585	0.0718	0.9272	0.9222