

**ANALYZING CONCEPTUAL GAINS IN INTRODUCTORY CALCULUS WITH
INTERACTIVELY-ENGAGED TEACHING STYLES**

by

Matthew Thomas

Copyright © Matthew Thomas 2013

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF MATHEMATICS

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2013

As members of the dissertation committee, we certify that we have read the dissertation prepared by Matthew Thomas, titled Analyzing Conceptual Gains in Introductory Calculus with Interactively-Engaged Teaching Styles and recommend that it be accepted as fulfilling the dissertation requirement for the degree of doctor of philosophy.

Guadalupe Lozano Date: **5/3/2013**

Deborah Hughes Hallett Date: **5/3/2013**

Jennifer Eli Date: **5/3/2013**

Nicole Kersting Date: **5/3/2013**

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Dissertation Director: **Guadalupe Lozano** Date: **5/3/2013**

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgment of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Matthew Thomas

Dedication

for Laura, Rick, Jeanne, and Kate

TABLE OF CONTENTS

LIST OF TABLES.....	8
LIST OF FIGURES.....	9
ABSTRACT.....	10
CHAPTER 1: INTRODUCTION.....	12
1 Introduction.....	12
2 Conceptual Knowledge.....	13
2.1 Concept Inventories.....	17
2.1.1 The Force Concept Inventory.....	18
2.1.2 The Calculus Concept Inventory.....	19
2.2 Conceptual Questions in Physics.....	20
2.3 Tools for Measuring and Encouraging Conceptual Gains.....	21
2.4 Measuring Gains.....	24
3 Interactive-Engagement and Interactive Instruction.....	25
3.1 Active and Interactive Learning.....	27
4 Connections between Conceptual Knowledge and Interactive Teaching.....	29
4.1 Interactive-Engagement and Concept Inventories.....	31
5 Additional Relevant Research.....	32
6 Counter-arguments to Interactive Learning Studies.....	34
7 Influence of Individual Variables.....	35
7.1 Gender.....	35
7.2 Prior Mathematics Exposure.....	37
8 Research Questions.....	37
9 Motivations and Anticipated Contributions of the Study.....	38
10 Design and Organization of the Study.....	42
10.1 Chapter 2: Analyzing Normalized Gain Scores and Interactively-Engaged Teaching.....	43
10.2 Chapter 3: Hierarchical Linear Modeling And Analysis of Individual-Level Predictors.....	44
10.3 Chapter 4: Comparing Gain Score Measures on the CCI.....	45
10.4 Chapter 5: Analysis of Classroom Instruction Using Item Response Theory.....	46
11 Limitations of Study.....	46
12 Implications For Practice.....	46
CHAPTER 2: ANALYZING NORMALIZED GAIN SCORES AND INTERACTIVELY-ENGAGED TEACHING.....	48
1 Introduction.....	48
1.1 Relevance to Prior Research and Research Questions.....	50
2 Methods.....	53
2.1 Coding Protocol.....	53
2.1.1 Types of Interaction Episodes.....	54
2.1.1.1 Initiator Independent Episodes: Developing Concepts.....	56
2.1.1.2 Public Student-Initiated Episodes.....	56
2.1.1.3 Public Instructor-Initiated Episodes.....	59

2.1.1.4 Private Work Times.....	62
2.1.1.5 Miscellaneous (Uncategorized) Interaction Count.....	63
2.2 Coding.....	64
3 Results.....	66
3.1 Counts, Sub-counts, and Correlations.....	66
3.2 Student Scores on CCI.....	72
3.3 Normalized Gain Scores on the CCI.....	73
3.4 Results of Comparing the Coding Protocol with CCI Scores.....	74
3.4.1 Main Hypotheses.....	74
3.4.2 Exploratory Analysis.....	77
4 Conclusions.....	80
CHAPTER 3: HIERARCHICAL LINEAR MODELING AND ANALYSIS OF INDIVIDUAL-LEVEL PREDICTORS.....	83
1 Individual Normalized Gains.....	86
2 Null Model.....	89
3 Pretest and Posttest Analysis.....	93
4 Analysis of Gender.....	95
5 Analysis of Previous Mathematics Experiences.....	102
5.1 HLM Using Prior Mathematics Course Groupings.....	108
5.2 Possible Correlation with SAT Scores.....	111
6 Conclusions.....	114
CHAPTER 4: COMPARING GAIN SCORE MEASURES ON THE CCI.....	116
1 Introduction.....	116
1.1 Relevance to Research Questions.....	117
2 Results of IRT Analysis and Implications.....	117
2.1 Types of IRT Models.....	118
2.1.1 Rasch Model.....	118
2.1.2 One Parameter Logistic Model.....	123
2.1.3 Two Parameter Logistic Model.....	128
2.1.4 Three Parameter Logistic Model.....	133
2.2 Comparing Models.....	137
2.3 Checking Assumptions.....	142
2.4 Computation of Gain Scores.....	146
2.5 Comparison of Normalized Gain and IRT.....	147
2.5.1 Advantages of Normalized Gain Scores.....	149
2.5.2 Advantages of IRT Gains.....	152
3 Conclusions.....	153
CHAPTER 5: ANALYSIS OF CLASSROOM INSTRUCTION USING ITEM RESPONSE THEORY.....	154
1 Relationship between IRT and IE Teaching.....	154
2 Instructor-level Results.....	155
3 Hierarchical Linear Models and Individual-level Analysis using IRT Gains.....	166
3.1 Null Model.....	166
3.2 Gender.....	166

3.3 Individual-level Predictors.....	167
4 Conclusions.....	171
CHAPTER 6: CONCLUSIONS AND IMPLICATIONS.....	173
1 General Comments.....	173
2 Implications for Future Research.....	173
3 Implications for Teaching.....	174
4 Future Directions.....	176
REFERENCES.....	178

LIST OF TABLES

Table 1: Types of Interactions Captured by the Coding Protocol.....	55
Table 2: Counts of Types of Interactions by Instructor.....	67
Table 3: Correlations Between Counts of Public Interactions Categories.....	69
Table 4: Correlations Between Counts of Public versus Private Episodes.....	70
Table 5: Regressions of Normalized Gains by Counts of Interactions.....	77
Table 6: Regression Results for Exploratory Analysis.....	79
Table 7: Results of HLM Null Model.....	90
Table 8: HLMs Predicting Individual Normalized Gains.....	99
Table 9: Regression Predicting Posttest Scores from Pretest Scores and Gender.....	100
Table 10: HLM Predicting Posttest Scores from Pretest Scores and Gender.....	101
Table 11: t-test Results from Previous Mathematics Course Questionnaire.....	103
Table 12: Normalized Gains for Students Based on Previous Mathematics Courses.....	106
Table 13: p-values for Pairwise t-test Between Mathematics Background Groups.....	106
Table 14: HLM Predicting Individual Normalized Gains.....	110
Table 15: Predicting Gains Based on Prior Mathematics Background and Placement Test	112
Table 16: Rasch Model Estimated Parameters.....	122
Table 17: 1PL Model Estimated Parameters.....	127
Table 18: 2PL Model Estimated Parameters.....	131
Table 19: 3PL Model Estimated Parameters.....	136
Table 20: Comparison of Pretest Rasch and Pretest 1PL Models.....	139
Table 21: Comparison of the Pretest 1PL Model and the Pretest 2PL Model.....	140
Table 22: Comparison of the Pretest 2PL Model and Pretest 3PL Model.....	140
Table 23: Comparison of the Posttest Rasch Model and Posttest 1PL Model.....	140
Table 24: Comparison of the Posttest 1PL Model to the Posttest 2PL Model.....	141
Table 25: Comparison of the Posttest 2PL Model with the Posttest 3PL Model.....	141
Table 26: Yen's Q3 Statistic.....	145
Table 27: Predicting Gains Scores Using All Interactions.....	162
Table 28: Predicting Gains by Number of Revisions Encouraged.....	165
Table 29: HLMs Predicting IRT Gains.....	167
Table 30: Pairwise Comparison of IRT Gains Based on Prior Mathematics Courses.....	168
Table 31: Mean IRT Gains of Students Based on Prior Mathematics Courses.....	169
Table 32: p-values of Pairwise t-test of IRT Gains Based on Prior Mathematics Courses	169
Table 33: HLM Predicting IRT Gains.....	170

LIST OF FIGURES

Figure 1: Normalized Gains versus Protocol Interactions.....	74
Figure 2: Normalized gains versus student initiated episodes.....	75
Figure 3: Normalized gains versus instructor initiated episodes.....	76
Figure 4: Normalized gains versus all interactions.....	78
Figure 5: Normalized gains versus number of revision encouragements.....	79
Figure 6: Comparison of classroom level and average individual normalized gains.....	89
Figure 7: Rasch Model for CCI Pretest.....	119
Figure 8: Rasch Model for CCI Posttest.....	120
Figure 9: 1PL Model for CCI Pretest.....	125
Figure 10: 1PL Model for CCI Posttest.....	126
Figure 11: 2PL Model for CCI Pretest.....	129
Figure 12: 2PL Model for CCI Posttest.....	130
Figure 13: 3PL Model for CCI Pretest.....	134
Figure 14: 3PL Model for CCI Posttest.....	135
Figure 15: Item difficulty estimates in sub-test versus full test.....	144
Figure 16: Normalized Gain vs. IRT Gain by Instructor.....	148
Figure 17: Individual Normalized Gains vs. IRT Gains.....	149
Figure 18: IRT gains versus CCI Pretest Scores.....	151
Figure 19: Normalized gains versus Student Initiated Episodes.....	156
Figure 20: IRT gains versus Student Initiated Episodes.....	157
Figure 21: Normalized gains versus Instructor Initiated Episodes.....	158
Figure 22: IRT gains versus Instructor Initiated Episodes.....	159
Figure 23: Normalized gains versus all interactions.....	160
Figure 24: IRT Gains versus All Interactions.....	161
Figure 25: Normalized gains versus number of Revision Encouragements.....	163
Figure 26: IRT Gains versus Number of Revision Encouragements.....	164

ABSTRACT

This dissertation examines the relationship between an instructional style called Interactive-Engagement (IE) and gains on a measure of conceptual knowledge called the Calculus Concept Inventory (CCI). The data comes from two semesters of introductory calculus courses (Fall 2010 and Spring 2011), consisting of a total of 482 students from the first semester and 5 instructors from the second semester.

The study involved the construction and development of a videocoding protocol to analyze the type of IE episodes which occurred during classes. The counts of these episodes were then studied along with student gains, measured in a number of different ways. These methods included a traditionally used measure of gain, called normalized gain, which is computed at the instructor level. Additionally, gains were further investigated by constructing hierarchical linear models (HLMs) which allowed us to consider individual student characteristics along with the measures of classroom interactivity. Another framework for computing ability estimates, called Item Response Theory (IRT), was used to compute gains, allowing us to determine whether the method of computing gains affected our conclusions.

The initial investigation using instructor-level gain scores indicated that the total number of interactions in a classroom and a particular type of interaction called “encouraging revisions” were significantly associated with normalized gain scores. When individual-level gain scores were considered, however, these instructor-level variables were no longer significantly associated with gains unless a variable indicating whether a student had taken calculus or precalculus in high school or in college was included in the

model. When IRT was used to create an alternative measure of gain, the IE variables were not significant predictors of gains, regardless of whether prior mathematics courses were included, suggesting that the method of calculating gain scores is relevant to our findings.

CHAPTER 1: INTRODUCTION

1 Introduction

On May 12th, 2011, an article in the *New York Times* described Carl Wieman's research on university introductory mechanics and quantum mechanics students (Carey, 2011). Wieman compared collaborative classrooms with active experimentation to those with no collaboration or experimentation, and showed that when introducing collaborative and experimental aspects to the large-lecture classroom, improvements were seen in student attendance, engagement, and learning (Deslauriers, Schelew, & Wieman, 2011; Wieman, 2007). The researchers found that learning among students in both quantum mechanics classes and second-semester introductory physics classes was greater in the interactive classes compared with traditional, lecture-based classes. In the experiment involving the introductory mechanics class, the researchers gave students a set of “clicker” (personal response system) questions designed and approved by the instructors of the course. The researchers found an effect size of 2.5, meaning that students in the interactive classroom outperformed their traditional lecture counterparts by 2.5 standard deviations. These gains are much larger than many other studies in education, where effect sizes are typically smaller than 1 (Deslauriers et al., 2011). Wieman's results build on the results of others such as Hake (1998), described further below, by suggesting that interactive classes in undergraduate physics have higher levels of conceptual gains than traditional lecture-based classes. Few similar studies in undergraduate mathematics have been conducted. Those that have, such as one by Epstein (2007), also described further below, have largely been less conclusive than those

in physics.

Wieman's results suggest a connection between conceptual gains and style of instruction in undergraduate physics classes which may also exist in mathematics classes. Further investigation of the factors influencing conceptual gains, including classroom interactivity level, is the goal of this study. Wieman's particular notion of interactive instruction, which included the use of clickers, is not used in this study. In his framework, a variety of components of the classroom, such as the inclusion of the clickers or the specific types of questions being used, may have contributed to student learning.

In the following sections, we describe some of the prior research which informs this project. Our goal is to better understand the connection between gains in students' conceptual knowledge and a particular framework for interactive teaching. We begin by discussing notions of what conceptual knowledge is and how it has been measured, concluding with a discussion of the particular instrument used in this study, the Calculus Concept Inventory (Epstein, 2007). We also discuss the issue of how gains are measured on this instrument. We then discuss various notions of interactive teaching, especially the framework used in this study, Interactive-Engagement (Hake, 1998a) . We also discuss the work that has been done to investigate the connections between interactive instruction and conceptual gains.

2 Conceptual Knowledge

In this study, we are particularly concerned with students' conceptual knowledge of the ideas in calculus. While we use a preexisting instrument to measure conceptual knowledge, we explore in this section some of the ways that others have defined,

encouraged, and measured conceptual knowledge.

Historically, there has been a division between the teaching of computational and conceptual material (Rittle-Johnson, Siegler, & Alibali, 2001). This contrast can be seen in the recent “math wars,” where proponents of traditional mathematics typically emphasize procedural fluency and proponents of reform-based (or standards-based) mathematics emphasize conceptual understanding (Schoenfeld, 2004), described as “sharply contrasting orientations” (A. G. Thompson, Philipp, T. Thompson, & Boyd, 1994, p. 1). Others, such as Wu (1999), have claimed that the dichotomy between the two types of knowledge does not really exist and that both are essential, while pointing out a need for more empirical research on the subject.

Rittle-Johnson et al. (2001) define *procedural knowledge* as “the ability to execute action sequences to solve problems” (p. 346). In contrast, *conceptual knowledge* is defined as “implicit or explicit understanding of the principles that govern a domain and of the interrelations between units of knowledge in a domain” (p. 346). For example, conceptual knowledge might be indicated by a student's understanding of the relationships between algebraic and graphical representations of functions, though a student might possess this knowledge without being able to verbalize it. One of the ways that conceptual knowledge can be demonstrated is by applying known principles or techniques in new situations. For example, the fact that students often do not recognize the same topic, such as optimization, in a different subject area or context gives credence to the claim that conceptual understanding is not yet obtained (Hughes Hallett, 2006, p. 4).

Others have defined necessary mathematical knowledge in such a way that it includes both conceptual knowledge and procedural skill. For example, when Ball et al. (2005) wrote about their efforts to find common ground between reform-oriented mathematics educators and traditionally-oriented mathematics educators, they included the need for students to have “proficiency with computational procedures” (p. 1056). Their definition of proficiency included both “computational fluency” and “understanding of the underlying mathematical ideas and principles” (p. 1056), producing terminology that encompasses what others might consider both conceptual and procedural knowledge.

Discussion around growth of conceptual and procedural knowledge often focuses on whether one type of knowledge develops before the other in a specific domain, with “concepts-first” and “procedures-first” proponents, each with evidence supporting their viewpoint (Rittle-Johnson et al., 2001, p. 347). Others have abandoned this distinction in favor of a framework that presumes that interactions between the two types of knowledge are more significant than either alone, and that either can precede the other (Rittle-Johnson et al., 2001). The two types of knowledge seem to be interrelated in which each builds upon the other (Rittle-Johnson & Alibali, 1999). This leads to an iterative process of building up knowledge (Rittle-Johnson et al., 2001). The relationship between procedural knowledge and conceptual knowledge is important to understand as we intend to study only one of these types of knowledge. If, for example, conceptual knowledge were highly dependent upon procedural knowledge, it would be more important to consider both types of knowledge in our study. If this were the case, instructors who fostered conceptual ideas would be limited by their students preexisting procedural

knowledge. Further investigation including measures of procedural knowledge may help us to better understand whether this relationship exists within our population.

Sfard, Neshet, Streefland, Cobb, and Mason (1998) considered the development of what they call computational and conceptual knowledge by considering their roles in discourse. Two types of discourse promote the respective type of thinking: computational discourse occurs when the primary topic of conversation is focused on calculation-based processes, but does not include discourse where the topic is specific instances of procedural manipulation of symbols. For example, presenting a solution to a given problem would not be considered computational discourse, while explaining how to do certain types of problems would be. Conceptual discourse is dialogue which focuses on the reasons for the calculations, and why they are done in the particular way that they are. This dialogue is heavily influenced by the sociomathematical norms of the classroom, since the expectations of justification may include varying levels of conceptual analysis depending on a particular classroom. These norms are influenced by the preferences and orientations of the instructor as well as the students (Thompson et al., 1994). In our study, we do not explicitly address the sociomathematical norms in the classroom, but we do consider the discourse in the classroom. The instructor's task of managing discussion with high conceptual demand is a task with its own set of challenges, and is discussed by Stein, Engle, Smith, and Hughes (2008). When teaching is done in this highly demanding way, a conceptual teaching style can lead to increased conceptual thinking by the students, though this level of demand is difficult to sustain (Stein & Lane, 1996).

The increasingly important role of conceptual understanding in mathematics has led

researchers to investigate ways to measure this type of understanding. In the following sections we describe the instrument used in our study, the Calculus Concept Inventory, and the history of concept inventories in general. We also describe how the conceptual knowledge measured by this instrument is related to procedural knowledge, and then discuss some of the ways that the encouragement of conceptual knowledge has been implemented and how gains in knowledge have been studied.

2.1 Concept Inventories

Conceptual understanding may be measured through instruments called concept inventories. Concept inventories are tests which are designed to measure the most foundational knowledge in a field (Epstein, 2007). The tests are typically given in a multiple choice format, and involve no computation. When given as a pretest and posttest, the instruments measure change in conceptual knowledge students experience during a course. Many studies of conceptual understanding in physics education use concept inventories (Hake, 1998a; Halloun, 1985; Libarkin, 2008; Malone, 2008; Rhoads & Roedel, 1999), and other disciplines have begun using them with increasing frequency. Concept inventories have been written for many subjects, such as statistics (Allen, 2006), precalculus (Carlson, Madison, & West, 2010; Carlson, Oehrtman, & Engelke, 2010), general biology (Garvin-Doxas, Klymkowsky, & Elrod, 2007; Smith et al., 2005), host-pathogen interactions (Marbach-Ad et al., 2009), natural selection (Anderson, Fisher, & Norman, 2002), general chemistry (Mulford & Robinson, 2002), the physics of waves (Rhoads & Roedel, 1999), astronomy (Prather & Brissenden, 2008; Prather, Rudolph, & Brissenden, 2009; Prather, Rudolph, Brissenden, & Schlingman, 2009; Rudolph, Prather,

Brissenden, Consiglio, & Gonzaga, 2010), and calculus (Epstein, 2007). This last one is the one used in this study. The results of studies using concept inventories have also influenced professional development (Marbach-Ad et al., 2010). In 2008, Libarkin documented a list of concept inventories in STEM fields, and this list did not include any in mathematics. The statistics concept inventory had been written at the time, though it was not included in the list. This suggests a lack of exposure of mathematics concept inventories, and their need for further investigation.

Many traditional tools for measuring knowledge focus on procedural knowledge. Rittle-Johnson et al. (2001) claim that even for the instruments which do measure conceptual knowledge, the pace of conceptual learning may be too gradual for pretest and posttest measures to be useful. Despite this, pretests and posttests can be and are used to measure conceptual understanding, though great care needs to be put into creating the measures (Epstein, 2007; Rittle-Johnson & Alibali, 1999). One particular challenge with measuring conceptual understanding is ensuring that the particular tasks have not already been completed, since no matter how novel the problem, reproducing a previously completed problem would be only demonstrating procedural knowledge (Rittle-Johnson et al., 2001).

2.1.1 The Force Concept Inventory

The first concept inventory to make a significant impact in the undergraduate education community was the Force Concept Inventory (FCI), written by Hestenes, Wells, & Swackhamer (1992). The FCI is a test in introductory mechanics which paved the way for analyzing student conceptual understanding of the basic ideas in a subject

area (Hake, 1998a, 2007; Hestenes & Wells, 1992; Hestenes et al., 1992).

The test was written to analyze students' thinking after realization that their commonsense beliefs were incompatible with Newtonian mechanics. Despite the fact that “the first impression of most physics professors is that the Inventory questions are too trivial to be informative,” (Hestenes et al., 1992, p. 2), students did poorly on the test. Of the 1,500 high-school students and over 500 university students who took the test, gains were quite low. High school students were reported to be learning 20%-23% of the previously unknown concepts, and college students at most 32% (Hestenes et al., 1992, p. 6).

Halloun and Hestenes (1985) define the knowledge required to successfully answer questions on the FCI as “common sense” ideas of mechanics, such as interacting forces. Frequently students think of interacting forces as a stronger force overpowering a weaker force, such as pushing a chair out-of-the-way, instead of an interaction according to Newton's third law (Hestenes et al., 1992). The results of their studies suggest that a large proportion of the students who do well by traditional measures of procedural skill in introductory mechanics courses have common-sense beliefs which are in direct contradiction with Newtonian mechanics.

2.1.2 The Calculus Concept Inventory

Drawing upon the FCI, Epstein wrote a concept inventory for introductory calculus in 2007. The 22 question multiple-choice test contains only non-computational questions, just as the FCI did. Similar to the FCI, the first pilot test for the Calculus Concept Inventory (CCI) was given to about 250 students at 6 schools in the spring of 2005, and

resulted in no gains anywhere, with pretest and posttest scores near the random guess level of 20% (Epstein, 2007). After modifying the test to make it significantly easier, “the conclusion was that if most faculty believe the test is trivial, we are probably about right” (Epstein, 2007, p. 168). It has since been given in at least 12 American universities and 1 in Finland.

Epstein states that “the Calculus Concept Inventory (CCI) is a test of conceptual understanding (and only that - there is no computation) of the most basic principles of differential calculus” (Epstein, 2007, p. 165). This description of conceptual understanding allows one to form a notion of the type of knowledge being assessed on a concept inventory by eliminating a certain class of questions which might be given to students, namely those which require procedural proficiency.

2.2 Conceptual Questions in Physics

Halloun and Hestenes' results showed that students, including those who received high grades and were able to successfully solve standard algorithmic problems, had poor understanding of the conceptual, or “common sense,” interpretations of the key ideas of Newtonian mechanics. Redish and Steinberg (1999) similarly found that students who were successfully able to answer standard, computationally oriented questions on topics such as the photoelectric effect still had ideas about photons which inhibited the way they thought about the nature of light. An example of a question probing conceptual knowledge in physics given by Redish and Steinberg (1999) is reproduced below.

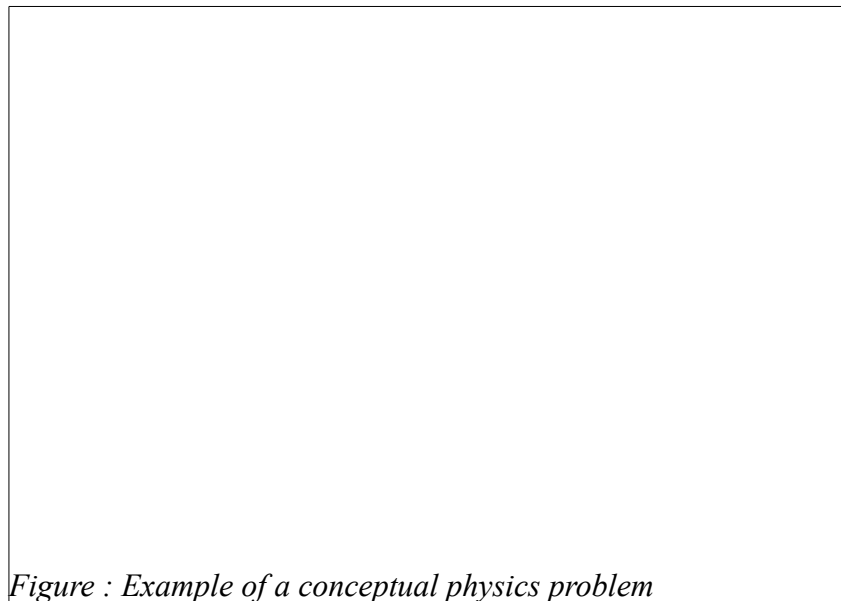


Figure : Example of a conceptual physics problem

This example shows that good conceptual questions elicit a different type of thinking than standard algorithmic problems. Students who are able to solve algorithmic problems may not necessarily understand the conceptual underpinnings of the subject areas. If developing conceptual understanding is a goal of a course, then non-standard questions like the ones used by Redish and Steinberg, the type present in concept inventories, can be useful tools for gaining insight into student thinking.

2.3 Tools for Measuring and Encouraging Conceptual Gains

While our study uses the Calculus Concept Inventory to measure conceptual understanding, many methods for evaluating and promoting conceptual understanding have been developed. One such method, ConcepTests, presents students with questions where “the purpose of the question is to have students start thinking about a new idea, to get their hands dirty, and to see how much they can figure out” (Cline & Lomen, 2009, p. 2). Another project, called the “Good Questions” project, aims to “raise the visibility of

key calculus concepts, promote a more active learning environment, support young instructors in their professional development in their early formative teaching experiences, and improve student learning” (Miller, Santana-Vega, & Terrell, 2006, p. 193). The project provides questions which have been used in college calculus classrooms to encourage active discussion of the content and lead to conceptual understanding by the students. These types of questions have been shown to improve student learning, though only when specifically used as a tool to encourage student discussion (Miller et al., 2006). An example of a “good question” is whether the statement “you were once exactly π feet tall” is true or false, where students may claim the statement to be false despite having a belief that height is a continuous function. The subject of introductory calculus is particularly good for exploring conceptual topics due to the presence of ideas such as limits and continuity algebraically and graphically (Koirala, 1997). Determining which questions are best, however, is a challenging task, as Miller et al. (2006) explain:

One might think for example, that a good question should be unambiguously clear, have a unique correct solution, and be framed with perfect mathematical precision. We found that the best questions, as measured by how frequently instructors chose to use them and on how successful questions were in stimulating class discussion, were those that had more than one interpretation, or had more than one or perhaps no solution. Questions like that were somehow more discussable. (p. 195)

The larger goal of all of these types of projects is to encourage classroom discussion, which is directly relevant to the larger goal of this study. These “good questions” are chosen in large part because they successfully encourage classroom discussions. The types of classroom discussions which may be encouraged by these questions are likely very similar to the types of classroom interactions which are of interest in this study. The

Good Questions project is based on Eric Mazur's Peer Instruction method (Mazur, 1997), developed for introductory physics classes (Terrell, 2003). Mazur's instructional ideas for interactive teaching were extended by Pilzer (2001) to include other physics and mathematics settings.

The Good Questions project has shown promise in raising test scores for instructors who choose to use the program (Miller et al., 2006). One of the results of this study, however, is that the questioning itself is not the most significant part of the project. What makes the most difference is the way in which the problems are used, specifically that they are used as a tool for motivating discussions amongst students. The type of classroom discussion which seems to provide the most benefit to students is of primary interest in this study. There have been studies done to investigate the questioning techniques of graduate teaching assistants, showing similar results (Roach, Roberson, Tsay, & Hauk, n.d.). When considering the discussions themselves, Cobb, Wood, and Yackel (1990) have found evidence that longer discussions occur among students who are discussing conceptually oriented topics.

The results of more quantitative studies, like those involving concept inventories, show positive results of interactive instruction on student learning, but they are not without controversy. In one series of articles, the merits of the FCI and interpretations of the results was debated, and a core component of the debate was how the results of the test should be used in practice. (Heller & Huffman, 1995; Hestenes & Halloun, 1995; Hestenes et al., 1992; Huffman & Heller, 1995).

The importance of considering conceptual understanding has been described in

general, and calculus is a particularly important subject for conceptual learning.

Conceptual understanding has been a key aspect of calculus reform (Hughes Hallett, 2006; Hughes Hallett, Robinson, & Lomen, 2005). The disconnect between high school and college mathematics classes, even high school calculus and college calculus, has been an active area of study, and differences in approach and style of thinking is often cited as a reason for the disconnect (Clark & Lovric, 2009; Long, Iatarola, & Conger, 2009; Mann, 1976; Panel, 1987; St. Jarre, 2008).

2.4 Measuring Gains

The method for measuring gains on a concept inventory has traditionally and almost exclusively been normalized gain, defined as:

$$\langle g \rangle = \frac{\textit{Posttest Score} - \textit{Pretest Score}}{\textit{Maximum Possible Score} - \textit{Pretest Score}} \quad (1)$$

In particular, this is the gain score that Epstein (2007) and Rhea (n.d.) used to report their findings on the Calculus Concept Inventory. The normalized gain score measures the fraction of previously unknown material that is learned throughout the course. For example, if a class average on a pretest were 40% and 70% on the posttest, the normalized gain would be $\langle g \rangle = 0.5$, meaning that class, on average, correctly answered half of the 60% of the material they answered incorrectly at the beginning of the semester.

Normalized gain is typically calculated using section averages of pretest and posttest scores, so each section of a course is assigned a single normalized gain score. Since many studies compare the effects of instructional practices on student learning, the effect of

interest is at the section level: normalized gains calculated at the section level allow one to analyze the effect of instructional practices on the entire class. One can also create an individual normalized gain score by using the pretest and posttest score for each student. The effect of computing individual normalized gains for each student has been investigated and compared to using section-level normalized gain scores (Bao, 2006; Coletta & Phillips, 2005). Bao found that the differences could largely be attributed to differences between classes where all students gained uniformly and those where the rank order of students changed. This change in rank order might occur in situations where an instructional style is particularly beneficial for students with lower initial ability levels, or if the instructional style is particularly effective for a subset of the population. The option of using individual normalized gains is explored in Chapter 3. The advantage of considering individual-level normalized gains is that instructor-level variables can be considered along with student-level variables such as demographics.

3 Interactive-Engagement and Interactive Instruction

One of the most influential uses of concept inventories has been in the comparison of instructional techniques. In this section, we discuss one framework for studying interactive instruction, Interactive-Engagement (IE), which is the framework used in this study. IE was defined by Hake (1998) as a collection of methods designed, at least in part, to promote conceptual understanding through “heads-on (always) and hands-on (usually)” (p. 1) activities which lend themselves to immediate feedback through discussion with peers and/or instructors. These classes are contrasted with traditional lecture (TL) classes, which are those classes that do not make use of IE techniques.

Instead, TL classes primarily rely on students passively listening to lectures, and measure knowledge with questions that are largely algorithmic (Hake, 1998a). As such, IE and TL classes can be differentiated both in terms of student activity during classes and in terms of the type of knowledge measured. For the purposes of this study, we consider IE classes to be those in which students are actively engaged with the material and interacting with the instructor, without regard to the specific type of knowledge the instructor is attempting to measure. IE teaching styles share features with Peer Instruction (Mazur, 1997) including ConcepTests (Pilzer, 2001), and pure discovery learning (Paris & Paris, 2001).

Epstein (2007) has similarly investigated IE instruction, investigating undergraduate mathematics instead of physics as Hake did. He builds on Hake's definition of an IE classroom in the context of studying introductory calculus classes and specifies that “in an IE class, students are actively engaged at all times, in developing concepts, developing strategies to solve problems of a non-routine kind, [and] testing solutions for sensibility as well as correctness” (Epstein, 2007, p. 166). Both characterizations share the critical component that student in-class work must receive real-time feedback from an instructor, other students, or a combination of these, and that such feedback must require sense-making and checks for consistency with other concepts already understood. It must also allow students to revise his/her conceptions accordingly. There are other studies which have investigated connections between student-instructor interactions and their effect on student success, such as a study by Deshler (2009). In her study, she considered “highly interactive” classes to be those with “the instructor keeping students on task, knowing

which students are present, showing an interest in their understanding of the material presented, and the presence of more academic discourse in the classroom” (p. 3).

Deshler's description of an interactive classroom has similarities with IE, though categories of discourse are constructed differently.

3.1 Active and Interactive Learning

While IE provides one way to frame instruction which directly involves students as an alternative to traditional lecturing, many other frameworks exist. Some, such as Inquiry-Based Learning, are broad categories which may incorporate many specific techniques, but have been shown to be positively associated with student learning in undergraduate mathematics (Laursen, Hassi, Kogan, Hunter, & Weston, 2011). Another framing for instruction which incorporates students into the learning process, for example, is “active learning.” Michael (2006) provides the following definitions:

Active Learning. The process of having students engage in some activity that forces them to reflect upon ideas and how they are using those ideas. Requiring students to regularly assess their own degree of understanding and skill at handling concepts or problems in a particular discipline. The attainment of knowledge by participating or contributing. The process of keeping students mentally, and often physically, active in their learning through activities that involve them in gathering information, thinking, and problem solving.

and

Student-Centered Instruction. Student-centered instruction [SCI] is an instructional approach in which students influence the content, activities, materials, and pace of learning. This learning model places the student (learner) in the center of the learning process. The instructor provides students with opportunities to learn independently and from one another and coaches them in the skills they need to do so effectively. The SCI approach includes such techniques as substituting active learning experiences for lectures, assigning open-ended problems and problems requiring critical or creative thinking that cannot be solved by following text examples, involving students in simulations and role plays, and using self-paced and/or cooperative (team-based) learning. Properly implemented SCI can lead to increased motivation to learn, greater

retention of knowledge, deeper understanding, and more positive attitudes towards the subject being taught. (p. 160)

Active learning and student-centered instruction are similar to IE in that both require students to engage with the material rather than passively receiving it. The example of active learning is provided to demonstrate how IE instruction might differ from other types of non-traditional instruction. IE instruction specifically requires that interactions must take place, whether the student is interacting with the instructor or other students, and so has some similarities with active learning techniques. Results of meta-analyses of university studies of active learning support the claim that active-learning supports student learning (Buck, 2005; Froyd, 2007). Since IE instruction and active learning have similarities, this suggests that IE instruction may also support student learning. While many studies have focused on active-learning in small classrooms, work in expanding this framework to large lectures in STEM is also being done (Smith et al., 2005). The active learning framework has also become a topic of undergraduate mathematics education discourse (Cline & Lomen, 2009; Hughes Hallett et al., 2005). While we aim to determine which aspects of IE instruction best encourage student gains, these results may not only be useful to those who are interested in IE instruction. It may also be of interest to those using other interactive instruction techniques, such as active learning.

It may be the case that the characteristics of IE physics classrooms that make them successful are not the same characteristics necessary for successful IE mathematics classrooms. The different nature of the two subjects, particularly the role of experimentation which is key to a physics classroom, may lead to differences in implementation. The classrooms in this study are relatively small compared to some other

studies. Mazur's (1997) Peer Instruction style was developed with large lecture classrooms of hundreds of students, while the classrooms in this study are all in the range of 30 to 35 students. The interactive instructional techniques which work in each type of classroom may inform the other, but may not be identical, and so class size may be another variable which should be considered in studying the effects of interactive teaching.

While interactive teaching, especially IE, may have many benefits to students, we are particularly interested in the potential effect that IE instruction may have on conceptual knowledge. In the next section we discuss the research, which has been done to investigate connections between interactive instruction and conceptual learning, and in particular we focus on the connections which have been found between IE instruction and gains on concept inventories.

4 Connections between Conceptual Knowledge and Interactive Teaching

The apparent connection between teaching style and conceptual learning documented by Hake (1998) has also been investigated in scientific disciplines such as astronomy (Prather, Rudolph, Brissenden, et al., 2009; Rudolph et al., 2010) and mathematics (Epstein, 2007; Rhea, n.d.). In biology, more research has been called for to determine whether a link between teaching style and conceptual learning exists (Libarkin, 2008), and biology-specific concept inventories are already being used to assess student understanding (Garvin-Doxas et al., 2007; Marbach-Ad et al., 2009). In this work we propose and carry out an approach to analyze IE classrooms which further fleshes out the hypothesized connection between this teaching style and student learning of concepts in

calculus.

Redish and Steinberg (1999) found that students in physics classes that they labeled as traditional instruction answered questions poorly on both a pretest and posttest, but students in classes with “modified instruction” improved substantially from the pretest to posttest. The “modified instruction” classes included those utilizing tutorials and “workshop physics,” both of which are used to teach calculus-based physics without lectures. These instructional methods are based on models from cognitive studies which suggest that students who develop understanding of material by building on previous understanding are more successful at developing conceptual understanding than those ignoring previously developed understanding (Redish, 1994). Additionally, they found that students in the “modified instruction” classes developed views of the subject area which they found more favorable, such as feeling a need for developing their own understanding and a need to evaluate material as opposed to taking statements by the instructor at face value.

Studies in mechanics which make use of a mode of instruction called “Peer Instruction” have similarly found that gains in conceptual knowledge are improved by allowing students to discuss questions with each other without any loss of procedural knowledge (Crouch & Mazur, 2001; Fagen & Crouch, 2002; Mazur, 1997).

In astronomy, a research group has conducted a national study investigating the effect of interactive teaching on conceptual knowledge (Prather, Rudolph, Brissenden, et al., 2009; Rudolph et al., 2010). By asking instructors to self-report their interactivity levels, they found that the highest gains were achieved by highly interactive classes,

demonstrating the possibility for highly interactive classes to encourage student learning. However, they also found that interactive instruction was not enough to guarantee high gains, as the range of gain scores was fairly large for both highly interactive and non-interactive classrooms. It is possible that these differences could be attributed to individual student differences.

4.1 Interactive-Engagement and Concept Inventories

One of the most well-known investigations of IE was Hake's (1998a) comparison of classrooms which compared scores on the FCI between classrooms with Interactive-Engagement (IE) methods with those which were described as “traditional lecture” (TL), including over 6,000 students. He found differences between the two types of classes of almost two standard deviations (Hake, 1998a, p. 65). Results from studies using the CCI have been less clear than those from the FCI. Rhea (n.d.) found large gains among highly IE classes, but Epstein (2007) found less conclusive results when including classrooms of more diverse interactivity levels. Epstein found very low gains for most classes between 0.08 and 0.20, with a few exceptions. These exceptional classes had normalized gains between 0.30 and 0.37, from instructors at the University of Texas at Austin, Oregon State, and St. Mary's College in Maryland. For comparison, Rhea (n.d.) found an average gain among 51 sections of 0.35, with 10 sections between 0.40 and 0.44. These high normalized gain classrooms were all taught using Interactively-Engaged styles, but some of the low normalized gain classrooms were also taught using these styles, suggesting that grouping classes by IE and TL may not be sufficient in college calculus. While these results are promising, statistically significant differences in gains in the CCI between IE

and TL classes have not been reported. A promising study conducted by Code, Kohler, Piccolo, and Maclean (2012), demonstrated higher performance on topics in a calculus class when a guest lecturer using IE methods taught particular topics to classes which was regularly being taught by a highly rated and successful lecturer, though statistically significant differences on traditional measures of student knowledge between the classes were not reported in the study. Similarly, Laursen et al. (2011) found that students in Inquiry-based calculus classrooms ranked higher in both course grades and in certain measures of affective constructs than students in non-inquiry-based classrooms, further supporting the hypothesis that interactive teaching is beneficial to students. It is, however, not clear whether conceptual understanding would necessarily follow same trend as procedural knowledge in relation to interactive instruction.

Though many types of interactive teaching techniques have been developed, the purpose of this section is not to illustrate differences between these techniques, but to illustrate that various models for interactive teaching seem to ultimately support learning in a variety of subjects. In this work we propose and carry out an approach to analyzing levels of Interactive-Engagement in classrooms which further refines the notion of an Interactively-Engaged class in introductory calculus by developing and analyzing categories of classroom interactions, and investigates the hypothesized connection between this teaching style and student learning of concepts in calculus.

5 Additional Relevant Research

The role of interaction has been explored in the K-12 setting as well. Hufferd-Ackles, Fuson, and Sherin (2004) studied the role of conversation with a Math-Talk Learning

Community framework and studied questioning, explaining mathematical thinking, sources of mathematical ideas, and responsibility for learning. This result supports a body of literature studying teaching and learning which studies what constitutes “good teaching.” Franke et al (2007) describe teaching as “relational,” so that “teachers, students, and subject matter can only be understood in relation to one another” (p. 227). Framing learning and teaching from this perspective implies that student-instructor interactions would be beneficial for student learning as it would provide a means for the student and teacher relationship. In a traditional lecture, where the interactions between instructor and student are always mediated by the content in a very specific way, namely a one-directional relationship, this relationship is very limited. It is reasonable to hypothesize that student-centered instruction would be more effective in encouraging learning.

Another way that K-12 research has influenced undergraduate research, specifically this study, is by demonstrating effective uses for video analysis (Kersting, Givvin, Sotelo, & Stigler, 2009; Kersting, Givvin, Thompson, Santagata, & Stigler, 2012). By analyzing the actions of teachers and students, these researchers have been able to connect instructional practices with student learning in a way that is more direct than questionnaires or interviews, and they provide a method for analyzing videos of college calculus students in this study. Mathematics learning is directly affected by the ways that teaching takes place in a classroom (Fennema & Franke, 1992; Hiebert & Grouws, 2007; Nye, Konstantopoulos, & Hedges, 2004). Video coding protocols have allowed researchers to analyze classrooms in a more in-depth manner than real-time observations

may allow. By re-watching segments of lessons, additional details may be noticed, and more people can reach consensus on what is occurring in the classroom. Video analysis has been useful both for providing teachers opportunities to respond to student work (Kersting et al., 2009, 2012; Norton, McCloskey, & Hudson, 2011) and to analyze teaching activities (Laursen et al., 2011).

Studies of Interactive-Engagement teaching techniques (Kost, Pollock, & Finkelstein, 2009; Miyake et al., 2010; Pollock & Finkelstein, 2007) in physics classrooms and Peer Instruction in physics classrooms (Lorenzo, Crouch, & Mazur, 2006) have demonstrated that these types of instructional strategies can reduce or remove gender gaps which existed at the beginning of the semester. This suggests that individual-level effects may be worth considering, which is done in Chapter 3.

6 Counter-arguments to Interactive Learning Studies

Kirschner, Sweller, and Clark (2006) claim that problem-based and inquiry-based learning are not effective for developing knowledge when compared to direct-instruction techniques. One of many responses to this paper was written by Hmelo-Silver, Golan Duncan, and Chinn (2007), who claim that Kirschner et al. misrepresent instructional styles by considering only extreme forms of inquiry-based learning, and reach incorrect conclusions because they do not consider the possibility of different levels of scaffolding. Kirschner et al. (2006) do not provide experimental evidence to support their claim, only providing general goals of information retrieval, and define learning as “a change in long term memory” (p. 75). As Schoenfeld (1995, 2004) describes, one trait of reform-based mathematics is a focus on problem solving instead of factual recall. It may be that

Kirschner et al. are not addressing these types of knowledge, as their paper does not directly address mathematical thinking.

Despite their arguments, Kirschner et al. (2006) claim “the advantage of guidance begins to recede only when learners have sufficiently high prior knowledge to provide 'internal' guidance” (p. 75).

7 Influence of Individual Variables

So far, we have focused on variables which are not specific to individual students, but instead apply to the instructional technique used in the classroom or are potentially common to all students. It is possible that individual characteristics play an important role in predicting gains on the CCI, or that IE instruction is particularly useful for certain students with particular characteristics.

7.1 Gender

One persistent problem in mathematics education is different success rates between male and female students. While these differences are decreasing, their cause is complex and still under active study (Hagedorn, Siadat, Fogel, Nora, & Pascarella, 1999; Reynolds & Conaway, 2003). Work by Seymour and Hewitt (1997) indicates that gender is a significant factor affecting whether students in STEM fields stay in their chosen major. Programs such as the Emerging Scholars Program have shown great promise in improving success rates of underrepresented minorities, including female students (Bonsangue & Drew, 1995; Fullilove & Treisman, 1990; Selvin, 1992; Treisman, 1992). Research in programs like the Emerging Scholars Programs has also illustrated that the type of active learning promoted among underrepresented minorities results in higher

grades among participants than non-participants (Moreno & Muller, 1999). The type of active learning involved in programs such as the Emerging Scholars Program is not identical to IE instruction, though it does share some characteristics, suggesting IE teaching methods are worth a closer look in connection to student characteristics such as gender.

In his construction and validation of the Statistics Concept Inventory, Allen (2006) found that neither gender nor ethnicity were correlated with student scores. This might be attributed to differences between mathematics and statistics or the types of students who are enrolled in each type of class. Calculus might be thought of as a continuation of the mathematics courses which have preceded it. College statistics, on the other hand, has many differences from many K-12 mathematics courses, and so it may be that the systematic differences in genders and ethnicities which exist in other mathematics courses are not present in introductory statistics. Additionally, the student population in an introductory statistics course is often different from that of a calculus course. Many students in an introductory statistics course are not STEM majors, and so the results involving STEM majors may not directly apply.

It may also be the case that the type of knowledge being measured on a mathematics concept inventory is sufficiently different from the type of knowledge traditionally measured in mathematics courses, so that gender differences which exist in traditional measures of knowledge in university courses are not as prominent as on concept inventories. This theory, however, seems unlikely, as gender differences have been observed on physics concept inventories, and IE teaching styles have been shown to

diminish the gender differences in physics courses (Kost et al., 2009; Lorenzo et al., 2006; Pollock & Finkelstein, 2007).

7.2 Prior Mathematics Exposure

Another individual-level variable of interest is prior exposure to mathematics courses, both at the high school and at the college level. In physics, Meltzer (2002) found that while students' incoming conceptual knowledge of physics did not significantly predict their normalized gains on the Force Concept Inventory, their incoming level of mathematical knowledge as measured by the ACT and a test of skills in algebra and trigonometry did.

Research has also been conducted in the effect of course-taking at the high school level on readiness for college-level mathematics courses. Adelman (2006), for example, found that taking mathematics courses beyond Algebra 2 significantly improved the probability of college completion. There is also a potential relationship between course background and gender, as male students are more likely to be ready for college level mathematics upon entering the university than female students (Long et al., 2009). There are certainly affective variables such as self-confidence related to course-preparation, though these were not considered for the current study. The effect of having taken precalculus or calculus courses before, and whether at the college or high-school level, are considered in Chapter 3.

8 Research Questions

The primary goal of this study is to investigate the following questions:

1. What individual-level and instructor-level (IE) factors affect gains in conceptual

knowledge as measured by the Calculus Concept Inventory?

2. How can the relationship between gains in conceptual knowledge as measured by the Calculus Concept Inventory and instructor-level (Interactive-Engagement) and individual-level (student demographic) variables be analyzed?

Each chapter addresses various aspects of these questions. In order to better understand the impact that instructor-level interactions may have on conceptual learning, the following additional questions are addressed:

3. What is a preliminary, non-self reporting model to characterize and quantify the level of IE in a classroom?
4. In the context of introductory undergraduate calculus classes, are conceptual gains, as measured by the CCI, correlated with IE teaching styles as measured by the scoring models developed? If so, how? If not, what does the data tell us?
5. Based on the videocoding protocol for measuring IE, which characteristics of IE classrooms are most correlated with gains on the CCI?
6. Do different methods for scoring conceptual knowledge gains correlate similarly with IE characteristics? How might we explain apparent differences and how important are they?

9 Motivations and Anticipated Contributions of the Study

The ideas of calculus are essential for students in many STEM fields. Many interested and talented students also leave the university or change to other fields of study during the time they are taking Calculus, and many of them cite poor teaching as a reason for leaving (Seymour & Hewitt, 1997). For this reason, calculus is an essential course for

studying student understanding, so that our results can have the most significant impact on students interested in STEM fields. While other studies have investigated students' abilities to solve traditional problems, the creation and validation of the Calculus Concept Inventory allows for tool to be used to measure student conceptual learning at this critical stage of students' college careers.

Our study also addresses potential self-reporting biases. Previous studies have relied on self-reporting levels of interactivity or have “binned” classrooms into two categories: IE or TL. Self-reporting introduces potential bias, as instructors may have difficulty accurately estimating their own level of interactivity, or may be inclined to report the way they think they teach or want to teach, which may not accurately reflect the way they actually teach (Cooney, 1985; Deshler, 2009; Raymond, 1997). We introduce a video-coding protocol to more objectively classify levels of IE, which we hope will better describe actual classrooms. In doing so, we can further determine which classroom activities most benefit students.

An additional benefit to the creation of a video-coding protocol is that it may be used as a systematic measure of instructional quality for other purposes. Classroom observations are an important part of graduate student and faculty careers, and many methods for measuring classroom performance are possible, though subjective. This protocol may provide a systematic way to observe and report instructional activities. The coding protocol developed in this study and preliminary results of the effects on student learning provide a springboard from which instructors can develop their own personal style of interactions to better support the particular students in their class. For example,

an instructor may find large-group discussions to be more effective than groupwork in small groups for a particular group of students based on their particular preferences. This study provides an initial attempt to quantify the effect which these interactions might have and provide a template on which future studies of this type may be based.

This study also provides a means for reliability not present in previous studies. When surveys are given to instructors to self-report whether their class is IE or TL, or the percentage of time spent in IE activities, only the instructor is determining this information. While having students also report on these factors is beneficial for understanding how students are interpreting the events in the classroom, the students may be perceiving the classroom very differently from the instructor. Differences between the instructor's evaluation of the classroom and students' evaluations of the classroom may provide more information about perceptions than about the realities of the classroom activities. By having two coders independently evaluate a subset of the videos and agree on over 80% of the decisions, a source of reliability has been introduced. It is assumed that additional individuals could then use the coding protocol and code videos reliably as well.

Research in physics education, particularly studies which have made use of the FCI, have revolutionized the way undergraduate mechanics is taught across the country (Savinainen & Scott, 2002). Undergraduate mathematics education is a relatively young research field which draws motivation and ideas from these other fields, but also is in need of determining whether results found in those fields transfer and, if so, how (Pilzer, 2001; Riegler, 2010). Our study contributes to the understanding of IE as it exists in

mathematics courses, which is not necessarily the way that IE is implemented in physics courses.

Physics and mathematics are different subject areas, which may affect the ways that interactions with the material are possible or beneficial to students. These differences exist both in the content and the teaching of the content. The content of an introductory mechanics course includes topics such as friction, which students have prior experience with. These experiences will lead students to have previously constructed conceptions, which may or may not be compatible with Newtonian mechanics. Regardless, these prior conceptions may provide a springboard for interactive instructors to engage with his or her students. Physical intuition may provide a context where, through various interactions, instructors may determine what misconceptions students have, and can ask probing questions allowing students to correct their misconceptions. While many of the tools of calculus are developed to solve problems in subjects like physics, the mathematical intuition relevant for a course like calculus includes an informal understanding of the behavior of functions. If a student has an intuition for the growth rates of different types of functions, an instructor can leverage this intuition to build concepts like L'Hopital's rule. In teaching, physics classes often include a lab component which can create a hands-on interaction with content. Mathematics classes with lab components are often computer oriented, and so the notion of interacting with the material in a lab setting may look different from other sciences. Interaction in a physics lab setting includes literal hands-on experience with objects to illustrate principles being studied. While hands-on activities can be developed in mathematics courses, there is no

hands-on analogue for studying L'Hopital's rule in the same way that there is for friction. Common sense notions in mechanics classes consist of understanding whether students' knowledge of classical Newtonian mechanics transfers to their ideas about the real world. Mathematics has different notions of “common sense” and a different relationship with the real world.

The novel method for measuring the level of Interactive-Engagement in a classroom introduced in this study may be used to draw preliminary conclusions about which types of interactions have the most influence on student learning. These conclusions should be taken as an indication of future directions that seem most promising to investigate using the expanded framework of IE. The methodological pieces to our work also demonstrate some of the ways that analysis of student-instructor interactions might be conducted in future studies, and their advantages over previously used methods. The statistical techniques used to analyze our data go beyond those previously used to analyze the CCI, for example. We discuss this further in the descriptions of the chapters below.

10 Design and Organization of the Study

This study addresses the first research question by combining the creation of a coding protocol for measuring levels of Interactive-Engagement in first-semester university calculus classes, student demographic information, and the measurement of student learning on an externally-validated instrument for conceptual understanding in calculus (the Calculus Concept Inventory). A total of 482 students volunteered to participate in the study, and 5 instructors volunteered to have 3 lessons each videotaped.

Analysis in all chapters utilizes the statistical software R (R Core Team, 2012) using

the `ggplot2` package (Wickham, 2009) for graphics. Analysis in Chapter 3 uses the `lme4` R package (Bates, Maechler, & Bolker, 2012) for construction of hierarchical linear models. Construction of item response theory models in Chapter 4 was done using a combination of the software tools BILOG-MG, a beta release of EQSIRT, and R package `ltm` (Rizopoulos, 2006), in addition to the software tools used in Chapter 4. Below, we briefly outline the content of each of the chapters.

10.1 Chapter 2: Analyzing Normalized Gain Scores and Interactively-Engaged Teaching

In this chapter, we discuss the development of the aforementioned coding protocol to better understand the types of interactions and activities which take place in introductory calculus classes, based on notions of IE present in the literature. A description of the coding protocol and its design are given, and counts of various types of episodes are provided for each of the classrooms investigated. These counts are used to classify levels of IE in classrooms. A type of episode, for example, might include a student suggesting an alternative method for solving a problem.

The coding of the videos is then used to find correlations with pretest-posttest gains on the CCI. Analysis of student gains is conducted using the most frequently used measure of gains in concept inventory studies, normalized gain. Correlations between counts of different types of interaction episodes and CCI normalized gain scores is given, providing preliminary evidence for possible connections between IE teaching styles and conceptual gains by students in introductory calculus classes.

This chapter sets the groundwork for addressing questions 3, 4, and 5 directly, and the

first two questions indirectly by constructing instructor-level variables to be analyzed and providing the traditional method of analysis to be compared with additional analysis methods.

10.2 Chapter 3: Hierarchical Linear Modeling And Analysis of Individual-Level Predictors

In Chapter 3, we discuss the benefits that a statistical technique called hierarchical linear modeling (HLM) provides in accounting for variables at both the instructor-level and the individual-level. Instructor-level variables are those which are derived from the IE coding protocol, and so would be the same for all students in a single classroom. The value of this variable is constant because the students are all in the same classroom, and not because each student had the same measure for the variable, which is accounted for by the HLM. This type of model allows for variance to be partitioned between the instructor-level and the individual-level which allows us to better explore the relationships between student and classroom characteristics and student learning. Hierarchical linear modeling is an extension of linear modeling which allows the nested nature of the data to be taken into account (Raudenbush & Bryk, 2002).

This chapter expands on the previous chapter by including individual gain scores where the previous chapter only considered class-averaged gain scores. In doing so, we find that nearly all the variance in the data lies at the student level, meaning that instructor-level variables are not able to explain differences in student gains alone. This fact was obscured in the previous chapter due to the averaging of class scores. This suggests that the relationships discovered in the previous chapter may have been

spurious, and require additional study. We do discover, however, that by including a variable which indicates previous mathematics courses taken, the IE variables from the previous chapter are statistically significant.

This chapter addresses both of the primary research questions by providing an additional statistical technique of HLM for analyzing the data, using both instructor-level and individual-level data.

10.3 Chapter 4: Comparing Gain Score Measures on the CCI

In Chapter 4, we address the issue of gain score measurement. The typically reported measure of gain on concept inventories, normalized gain, is the gain score reported in Chapter 2 and Chapter 3. Another method for measuring ability coming from the field of psychometrics, called Item Response Theory (IRT), is discussed and used to analyze the CCI. Item Response Theory is not frequently used in concept inventory analysis, and an IRT analysis of the CCI has not been published. This methodology provides an additional way to measure student gains using the same data, and since neither method of measuring gains is objectively better than the other, it allows us to consider whether an additional method for measuring gains results in a different interpretation of the data.

We create and utilize IRT gain scores following Wallace and Bailey (2010), and use these scores in two ways. In this chapter, the IRT scores allow us to analyze the questions on the CCI to determine what new information we can find about the instrument and the students taking the test. The second way we use the IRT gains is discussed in Chapter 5, where we return to the study of IE classrooms and their connection to conceptual gains. Chapter 4 specifically addresses research question 6.

10.4 Chapter 5: Analysis of Classroom Instruction Using Item Response Theory

In this chapter, we use the IRT gains scores developed in the previous chapter along with the normalized gain scores discussed in Chapters 2 and 3 to determine whether the IRT gains perspective provides additional insight to the hypothesized connection between IE classrooms and conceptual learning. We consider how this additional method for computing gains may affect the proposed possible relationship between conceptual gains on the CCI and IE methods of instruction.

11 Limitations of Study

Though the study includes 482 students, the number of instructors videotaped (5) imposes certain limitations. Since only a subset of all the instructors were videotaped, we have information about the instructor's IE practices for 130 of the students. The five videotaped instructors represent a variety of instructional styles, but the sample is not large enough to make generalizations beyond the data in the sample.

Additionally, the videotaping was conducted the semester after the CCI was given to students, which introduces an additional possible source of error. We invited instructors who taught the course both semesters to participate, and all instructors in the study were experienced, and had developed their own instructional style. Instructors were asked about differences in instructional approach and classroom environment, and all claimed that these were very similar between the two semesters, and that no intentional changes in style occurred.

12 Implications For Practice

This study has potential applications to practice to those who teach university classes

or train others to teach. For instructors, the results of this study may provide insight or ideas for ways that they might modify instruction to better encourage conceptual learning among their students. For trainers of instructors, the coding protocol used to classify and quantify interactions may be useful in professional development programs to provide more objective feedback to instructors. It may be useful to have instructors consider how interactive they consider their own teaching to be, and then provide them the opportunity to compare their impressions of their teaching with measures based on the protocol. The opportunity for reflection is one which can be very useful for instructors, and a protocol such as this may be useful for providers of professional development to new instructors, particularly those new to teaching undergraduate mathematics such as graduate teaching assistants or new faculty. Additionally, information about student perceptions of instructor engagement and interactions, either by administering polls or by using end-of-semester evaluations, may provide a third piece of information along with instructor perception and the protocol developed in this study to better understand the realities of instructional practices. These data may be useful for instructor evaluations, as well as in introducing new faculty to the particular culture of the department.

CHAPTER 2: ANALYZING NORMALIZED GAIN SCORES AND INTERACTIVELY-ENGAGED TEACHING

1 Introduction

Reform in mathematics education has encouraged changes in content in K-12 education from the 1989 NCTM Standards (Franke et al., 2007; Hiebert, 2003; Hufferd-Ackles et al., 2004; Monson, 2011) to the recent Common Core State Standards (Carmichael, Martino, Porter-Magee, & Wilson, 2010), and in undergraduate mathematics education through calculus reform (Hughes Hallett, 2006; Schoenfeld, 1995). These changes have demonstrated that encouraging active learning techniques such as groupwork and discussion of content are beneficial to students (Cohen, 1994; Hufferd-Ackles et al., 2004; Schoenfeld, 2004). In this chapter, we propose ways in which interactions in college calculus classrooms may be studied, and the effects that these interactions may have on students' conceptual knowledge of the content.

Past president of the Mathematical Association of America David Bressoud (2011) stated that mathematicians commonly believe lecture is an effective teaching tool, though there is evidence to the contrary. The Higher Education Research Institute at the University of California in Los Angeles reported that “sixty-three percent of STEM professors said they used “extensive lecturing” in all or most of their classes,” while “about 37 percent of faculty in other fields said they did so” (Berrett, 2012b, p. 1), suggesting that lecture is still the primary means by which mathematics is communicated in universities. While members of mathematics departments were included in the poll, the specific instructional practices of mathematics professors were not reported. Current

research includes investigations of instructional ways that STEM content may be conveyed more effectively. Bressoud provides examples from physics, including Wieman's work, mentioned in Chapter 1, that show interactive classrooms may result in larger student knowledge gains, and suggests that the relationship between teaching and learning methods observed in undergraduate physics classrooms may also exist in undergraduate mathematics classrooms, though possibly in different ways.

One effect of current instructional styles is that it drives many talented students out of STEM fields into other fields or out of the university (Seymour & Hewitt, 1997). While students leave STEM fields for many reasons besides instructional style, the number of students who leave STEM fields is a concern. The President's Council of Advisors on Science and Technology recently reported that “less than 40 percent of those who enter college intending to be STEM majors complete a degree in one of those fields” (Berrett, 2012b, p. 1-2). With so few interested students completing STEM majors, it is worth considering whether instructional changes can be implemented which encourage more students to be successful in mathematics.

The *New York Times* article described in Chapter 1 includes one critique of Wieman's research: it is difficult to determine which changes in instruction accounted for the changes in student success. When observing a highly interactive classroom and a more traditional lecture-based classroom, one notices many differences between the instructional styles. Teaching techniques such as Interactively-Engaged teaching (Hake, 1998a), Just-in-Time teaching (Mazur & Watkins, 2010; Novak, Gavrin, Christian, & Patterson, 1999), and Peer Instruction (Crouch, Watkins, & Fagen, 2007; Mazur, 1997)

are all examples of methods for “flipping the classroom” (Berrett, 2012a), where students do not passively receive information, but are instead actively involved in the learning process. An instructor may include some or all of the elements of any one of these interactive instructional styles into his or her teaching, and so it may be useful to think of interactive teaching as a collection of possible techniques instead of a single choice which an instructor can follow completely or not at all.

The novel method for measuring the level of Interactive-Engagement in a classroom introduced in this chapter allows us to draw some preliminary conclusions about types of interactions that may be more (or less) associated with student gains.

1.1 Relevance to Prior Research and Research Questions

One of the goals of this chapter is to operationalize the concept of an Interactively-Engaged (IE) classroom in a way that Interactive-Engagement can be quantitatively measured. This characterization allows us to gather observational data which is then further analyzed in later chapters. A second goal of this chapter is to explore potential correlations between our measure(s) of Interactive-Engagement (constructed from the data) and gains on the CCI.

As mentioned in Chapter 1, previous studies which consider correlation between concept inventory scores and interactive instruction have relied on instructor (and sometimes student) self-reporting to quantify levels of IE in classrooms. For example, a study by Prather et al. (2009) relied on instructor self-reporting of interactivity levels, where questions were designed to determine how frequently “interactive learning strategies” (p. 322) were implemented, and how often students made predictions by

themselves or were asked questions during class. Rhea's (n.d.) study relied on student and instructor reporting of interactivity levels. Students were asked:

If an interactive classroom is one in which students actively work on underlying concepts and problems during the class and receive feedback from the instructor or other students on their work in class, how would you describe your class this semester? [Very Interactive; Interactive; Somewhat Interactive; Not Interactive] (p. 2)

and

On average, about what percent of your time in class would you say was spent with you working on problems and receiving feedback from your instructor and/or your classmates? [76-100%; 51-75%; 26-50%; 1-25%; 0%] (p. 2)

Consistency would require that a single observer would quantify levels of IE in different classrooms in the same way, but since the coders (the instructors themselves) are only coding their own classrooms, consistency cannot be assumed. Each instructor needs to interpret terms like “active,” “underlying concepts,” and “feedback” according to their own ideas of these, and one cannot assume that different instructors understand these terms in the same way. For example, some students or instructors may consider having students raise questions in class to be interactive, while some might require a higher degree of input, such as students presenting solutions to the entire class or generating solutions in groupwork, to consider a lesson interactive.

Reliability requires that different individuals would quantify IE levels in a single classroom in the same way. Since each instructor is only scored once, reliability cannot be assumed. When instructors self-report their own interactivity levels without verification by other observers, neither reliability nor consistency is achieved. By creating a coding protocol, we are able to achieve consistency, and by having more than one individual view and code videos independently, we are able to quantify the level of

reliability achieved.

An additional limitation to previous studies such as Hake's (1998) and Epstein's (2007) is that classes are considered to be either “Interactively-Engaged” or “traditional lecture,” without allowing for the possibility that classes may lie somewhere between these two categories, or that IE might be characterized in multiple ways. Our study builds on previous work by introducing a new way to describe and quantify the level of IE in a classroom. Previous studies have also considered an entire collection of instructional techniques to be included in an IE classroom, suggesting that a classroom is either an IE classroom or it is not (as done by Hake (1998a)). Others have considered some percentage of time spent in all types of IE activities (as done by Rhea (n.d.)). In our study, we attempt to deconstruct the notion of an IE classroom so that specific types of interactions can be identified. For example, in the protocol described below, we divide interactions into various types depending on the initiator of the interaction and the nature of the content discussed, such as whether the conversation was designed to extend the conversation beyond the immediate problem or was intended to suggest an alternative solution strategy. It is possible that different types of interactions are effective to different degrees in encouraging learning, so differentiating these types of interactions allows us to investigate potentially different effects on the learning of concepts.

Specifically, the construction of the protocol addresses the third research question, “What is a preliminary, non-self reporting model to characterize and quantify the level of IE in a classroom?”. The scores arising from the protocol allow us to begin to address research question one, “What individual-level and instructor-level (IE) factors affect

gains in conceptual knowledge as measured by the Calculus Concept Inventory?” by finding correlation between these instructor-level variables and gains on the CCI. We are then also able to address question five, “Based on the videocoding protocol for measuring IE, which characteristics of IE classrooms are most correlated with gains on the CCI?” by considering correlations between specific types of IE activities parsed out by the coding protocol and CCI gains. This sets the stage for addressing research question two, “How can the relationship between gains in conceptual knowledge as measured by the Calculus Concept Inventory and instructor-level (Interactive-Engagement) and individual-level (student demographic) variables be analyzed?”, which is continued in subsequent chapters.

2 Methods

2.1 Coding Protocol

All students taking introductory calculus in the fall semester of 2010 took the CCI as a pretest and posttest. In order to determine whether specific instructional practices were associated with gains in conceptual knowledge as measured by the CCI, instructors who were again teaching introductory calculus in the spring semester of 2011 were invited to participate in the study. These instructors were targeted because we would have CCI results from their students the previous semester along with observation of their instructional styles. Of the ten instructors who were teaching introductory calculus again, five agreed to participate in the study. Instructor participation involved agreeing to be videotaped in the classroom three times during the semester for the entire class.

The development of a coding protocol often involves constructing a skeleton of what

the coders expect to see in the videos, and how they intend to code the videos. In our case, we developed a set of interaction types we expected to see with descriptions of what would constitute each type of interaction. We then used three of the videos to test this preliminary protocol and develop it further. As examples emerged, the descriptions of the interactions were refined, and key examples were developed. Additional categories which were not anticipated were also included in the protocol. The remaining 12 videos were coded using the protocol.

All instructors were teaching from a common syllabus and were teaching towards a common final exam. Though instructors were videotaped on the same days, the lessons videotaped were not always over the same material. Despite this, topic overlap existed, so comparisons of the videos were reasonable. If topics had been very different, it could have biased the results, since some topics might lend themselves to IE teaching styles more naturally, so instructors who happened to be teaching those topics would appear to have higher levels of IE.

2.1.1 Types of Interaction Episodes

Table 1 illustrates all the types of classroom activities captured by the protocol, and descriptions and examples of each follow. Episodes are first categorized as being public or private, and then are categorized by the initiator of the episode. Public episodes are then further divided by the type of interactions which takes place. When private work time occurs, the number of interactions is counted and the initiator is noted. The total amount of time spent working is also recorded. This time is counted separately depending on whether the private work is groupwork or individual work. One reason for separating

private work from public work is that each may contribute to student gains in different ways, and the literature does not currently distinguish between these types of episodes in an IE classroom. By dividing episodes in this way, we can begin to determine whether there is evidence that public or private episodes encourage gains more than the other.

Table 1: Types of Interactions Captured by the Coding Protocol

	Initiator-independent	Student-Initiated	Instructor-Initiated
Public Episodes	Developing Concepts	Developing Strategies	Promotes sensemaking
		Sensemaking	Promotes checks/connections to previous material/extensions beyond current material
		Checking for correctness	Encourages revisions from students
			Check procedures for sensemaking
			Presentation of Problems Worked on by Students
Private Episodes	Groupwork time	Private, Student-Initiated episodes	Private, Instructor-Initiated Episodes

The initiator of an episode is the person who introduces the content of a conversation. In public conversations, this is typically very clear, as the instructor typically initiates episodes unless a student specifically asks a question or proposes an idea. An instructor may invite questions, for example saying “Any questions?”, which was interpreted as an invitation for students to initiate episodes. In private conversations, an instructor may walk around a classroom asking students whether they have any questions. This does not, in itself, constitute an initiation to an episode. If the instructor asks a specific question about the content, then the instructor has initiated the episode. If a student raises their

hand to ask a question or if the student responds to a general question like “do you have any questions?”, the student is considered to have initiated the episode.

2.1.1.1 Initiator Independent Episodes: Developing Concepts

In Hake's and Epstein's description of IE classrooms, the notion of developing concepts is described. We interpret this to mean that students should be involved in the construction of fundamental concepts in the course. The wording of the IRB stipulated that, to protect the anonymity of the students, student comments would not be used, so each of these examples is constructed as a fictional example which mirrors the type of interactions which received credit in that category.

Developing Concepts

Developing concepts is the only category which is not required to be specifically initiated by the instructor or by a student. It may be initiated by either, but consists of a sustained discussion on the conceptual content on a topic. Student contribution to the conversation, however, is required for an episode to be considered as developing a concept.

Example: Examples of developing concepts include sustained discussion about the ideas behind Riemann sums or L'Hopital's rule. For example, an instructor might develop the idea of L'Hopital's rule by appealing to notions of derivatives and rate of change to motivate the statement of the rule.

2.1.1.2 Public Student-Initiated Episodes

The student initiated episodes consist of developing strategies, sensemaking, and checking for correctness. These types of episodes are all derived from the descriptions of

IE classrooms given by Hake and Epstein, and the specific requirements to be counted in these categories were developed as the first set of three videos was coded. These episodes further develop our operational definition of an IE classroom. Classrooms where more of these episodes take place are considered to contain more Interactive-Engagement. By keeping types of episodes separate, we can also determine which types of episodes occur in high IE classes.

We first discuss episodes which are public in nature. These episodes typically take place when a student asks a question or makes a suggestion during class by raising their hand. If the student makes a suggestion which extends the conversation beyond the scope of the current conversation, this is a new student-initiated episode, as opposed to a continuation of the occurring episode. Student-episodes can include incomplete attempts, such as an incompletely formed question or suggestion. For student-initiated episodes, a student attempting to contribute to the discussion is the key factor in determining the student as the initiator.

Developing Strategies

A student initiated interaction which received credit for developing strategies is one in which a student suggests or asks a question about how to solve a problem. This may be a suggestion or question specific to the problem at hand or about a class of problems. Examples of developing strategies are suggesting a new step in a problem or asking whether a different solution path would be successful.

Example:

Instructor: ... And so by L'Hopital's rule, the limit is $\frac{1}{2}$. Yes, [student]?

Student: What if we had gotten 0 over 0 again? Could we just use L'Hopital's rule again?

At this point, the student has suggested a strategy, so the instructor might respond by providing an answer or delaying a response until later in the class. The student's suggestion for a strategy is all that is required for credit in this category, provided the instructor acknowledges the comment and responds in some way, so that there is an interaction.

Sensemaking

Sensemaking episodes are those student-initiated episodes in which a student makes a comment or raises a question about interpreting content in the course. This may involve interpreting answers, units, magnitudes, or signs of answers in the work being discussed.

Example:

Instructor: ... So the derivative of x is -3 . Questions?

Student: Would that mean the particle is moving to the left?

Checking for Correctness

These episodes are those in which a student makes a comment which corrects or asks about the correctness of a solution or step in a solution process. Examples of these types of episodes are when a student asks why a particular step in a process was justified, or when they point out a mistake an instructor may have made, whether the correction is justified or not, and whether the instructor's "mistake" was made intentionally or not.

Example:

Instructor: We then plug in the values and get 27 as the final position.

Student: Shouldn't that be negative 27?

Instructor: Right, yes it should.

2.1.1.3 Public Instructor-Initiated Episodes

Instructor initiated episodes are those in which the instructor specifically asks a question or begins an interaction where the instructor has determined the topic of the conversation. The interaction types in this category are: promotes sensemaking, promotes checks/connections to previous material/extensions beyond current material, encourages revisions from students, check procedures for sense-making, and presentation of problems worked on by students.

Promotes sensemaking

Sensemaking episodes made by instructors follow the same basic description as those which are student initiated. These typically take the form of an instructor making a suggestion about how to think about a problem or type of problem. It may be drawing attention to notation, perhaps noting where a parameter is being used in a new way, or when approximation is being used in place of an exact solution, such as estimating Riemann sums.

Example:

Instructor: So what does this t tell us?

Student: The places where we're summing the rectangles.

Instructor: Right – and remember this is the same t that was used for time in the original problem.

Promotes checks/connections to previous material/extensions beyond current

material

These episodes are those where the instructor extends the discussion outside of the immediate context by asking students to check some piece of the work, or connecting the immediate material to material that has already been covered or will be covered in the future. This may be a connection to either content material or to comments made by students.

Example: (Riemann sums)

Instructor: What happens as n gets larger?

Student: The difference between the left and right hand sums gets smaller.

Instructor: Right, and that's the rectangles that we drew earlier getting smaller and smaller.

Note: The exchange above already constitutes an interaction, whether students respond to the instructor again or not. The students do not need to acknowledge the instructor's reply or follow up.

Encourages Revisions from Students

These episodes require the instructor to explicitly suggest a revision from the students in the class. This may be a revision of work the instructor has written himself/herself, or may be a suggestion to improve upon or correct work presented by a student. It may also take the form of checking a student's answer by posing it to the classes. What is key here is not that a revision occurs, but that the type of thinking required to make revisions is engaged in by the students.

Example 1:

Instructor: So what's the derivative of $\sin(x^2)$?

Student: $\cos(x^2)$.

Instructor: What do you guys think?

Student: No, you have to use the chain rule.

Note: This type of episode can receive credit whether the suggested correction is correct or incorrect.

Example 2:

Instructor: So I need to differentiate $x = \cos(\theta)$ with respect to t , so I get

$$\frac{dx}{dt} = -\sin(\theta)$$

What do we think?

Student: You still need to have a $d\theta/dt$ because of the chain rule.

Instructor: Ok, good.

Check procedures for sense-making

These episodes are those in which an instructor suggests or asks a question about checking the steps of a solution process to make sense. This might include asking why a particular step was done as opposed to a different step. These episodes are distinct from the sensemaking episodes in that sensemaking episodes are about types of problems, or how to think about the content in the course. Checking procedures for sense-making is a category which captures interactions focused on the specific details of solving a specific problem, such as determining why a particular step of a solution process is justified.

Example:

Instructor: So how do we want to start this problem?

Student: We need to draw a picture so that we can relate the variables to each other.

Instructor: Ok, what kind of picture are we going to get, then?

Presentation of Problems Worked on by Students

This category was created to capture specific types of interactions not suggested by the IE literature by Hake (1998a, 1998b) or Epstein (2007). These were instances where an instructor presented the solution to a problem on the board after students had worked on the problem either individually or in groups, and had completed work on the problem. These are not interactions in the typical sense as they do not necessarily include a verbal exchange between students and instructor, but are interactive in that the instructor is providing direct and immediate feedback to students on their own work immediately after completion. These episodes provide a unique type of interaction from the others in that students are able to consider their own work in a public setting, thus interacting with the material in a more public way than in small groups or individual work. Since the instructor is presenting work to all students at the same time, the potential benefits of public interactions may be gleaned, unlike private interactions which only benefit one or a few students at a time.

2.1.1.4 Private Work Times

Since groupwork allows students to provide each other with immediate feedback and individual work time provides students opportunities to engage with content, the amount of time students spent on each type of work was recorded. The amount of time devoted to groupwork between the five instructors varied greatly, and has the potential to be another characteristic of an IE classroom. The amount of time for private work was only

considered if the private work lasted at least two minutes. Shorter episodes did not allow students to engage with each other, or the questions beginning the private work were not of sufficient difficulty to encourage in-depth, conceptual conversations.

In addition to time being provided for groupwork, many of the private work episodes occurring in classes also included instructor-student interactions as the instructor circulated the room. The number of these episodes was recorded, divided by who initiated the episode. Student initiated episodes were those in which a student raised their hand to ask the instructor a question or responded to an instructor's content-irrelevant question like "how are you doing?". While we were not able to analyze the specific details of the conversation, it was clear from the instructor's behavior whether he or she was asking a pointed question or just inviting questions. Instructor-initiated episodes are those in which the instructor asked a specific question of a student, instead of a general question which only invited conversation but did not initiate discussion of the content. This followed the same pattern as determining the initiator for public interactions where an instructor asking "Any questions?" was not considered to initiate an instructor-initiated episode.

2.1.1.5 Miscellaneous (Uncategorized) Interaction Count

In addition to the interactions coded using the protocol, a category was created to capture the interactions which did not fall into any of the other predefined categories. These included interactions where the topic was precalculus material or may have not qualified as any other particular type of interaction.

2.2 Coding

After the protocol was created, one video from each instructor was coded by each of two separate researchers. This coding was done independently, and a master sheet was created to resolve any disagreements. Reliability was greater than 80% for each of the coders on each decision, for each video.

The process of coding a video consisted of counting the number of occurrences of each type of episode. To facilitate the computation of reliability, each lesson was broken into 10-minute-long segments. Since only the count of interaction types was considered, this segmentation played no role in the actual coding, as segmenting into any other time length would result in the same counts. The segmentation did, however, allow for ease in computing reliability between two video coders. If no episodes of a given type occurred during a 10 minute segment, this was considered a single decision. For example, if the first 10 minutes included two “developing concepts” episodes in the master code, and there were no other “developing concepts interactions” in the entire 50-minute lesson, there would be a total of 6 decisions to be made: the first 2 interactions would each be a decision, and the decision that for each 10-minute segment after the first 10 there was no interaction. Therefore, each individual would receive a score out of 6, and this percentage would be their reliability for that type of episode in that lesson.

The length of an episode was determined by the framing of the question or comment which initiates the episode. For example, an instructor might ask “what is the value of x in this problem?”. In this case, the question may mark the beginning of the episode, and the end of the episode would come about when the value of x is determined. If the

instructor instead asked “how would we set this problem up?”, the episode would be considered to conclude when the setup for the problem has been addressed. Though not frequent, this may allow for a single episode to include multiple exchanges and/or multiple students.

Only non-routine problems were considered admissible for the code. For the purposes of this study, routine problems are those which are completely procedural; they require no interpretation and are algorithmic in nature, such as asking students to find the derivatives of a list of functions. In the classrooms observed, these types of problems were very uncommon, only occurring a few times. In the lessons, most problems were presented with some type of context or were building towards some type of discussion about underlying concepts. For example, any related rates problems were considered to be non-routine because they allowed for interpretation of the solution method, such as determining how to model the situation or how to interpret a solution in real-world terms. A problem involving a conical sand pile might include a conversation about the shape of a sand pile, or the sign of the rate of change of the radius with respect to time might be interpreted as meaning that the radius was increasing at a particular time.

In order to be considered for the specific types of episodes, the content of the interaction needed to be calculus-based. An additional category of “Miscellaneous episodes” was created to capture interactions which either did not fit into one of the created categories or were on content other than calculus, such as evaluating trigonometric functions. Additionally, questions needed to access students' knowledge as opposed to students' perception of their knowledge of the material. Simply asking choral-

response questions, where the answer was clear from the question was considered inadmissible. Assessing students' perceptions occurred frequently when an instructor asked “does that make sense?”. An answer to this question does not provide the instructor with any information of students' understanding, only whether they think they understand, or are unwilling to admit that they do not understand. Similarly, choral response questions almost never provide substantial information to an instructor. These typically only assessed student perception of understanding, and never provided opportunities for discussion to continue. If a choral response question did lead to a substantial conversation, this conversation was eligible to be counted as an interaction episode.

3 Results

3.1 Counts, Sub-counts, and Correlations

The counts of types of episodes are given in Table 2 along with the normalized gain scores.

Table 2: Counts of Types of Interactions by Instructor

Instructor		A	B	C	D	E
	All Instructor Initiated Episodes	42	29	25	23	30
	All Student Initiated Episodes	10	4	8	16	24
	Sum of Instructor and Student Initiated Episodes	52	33	33	39	54
	Developing Concepts	3	1	0	0	0
Student Work-Time, Including Private Episodes	Groupwork time (seconds)	1872	1249	0	865	0
	Individual Work Time (seconds)	0	504	0	2939	0
	Total Work Time (seconds)	1872	1753	0	3804	0
	Instructor initiated private episode	11	18	0	0	0
	Student-initiated private episodes	16	14	0	38	0
Instructor-Initiated Public Episodes	Promotes Checks	6	7	6	5	2
	Encourages revisions from students	4	8	0	5	6
	Promotes sense making	9	5	1	3	3
	Feedback on questions answered by students	22	9	14	2	18
	Problem presented which students have worked on	1	0	4	8	1
Student-Initiated Public Episodes	Student initiated developing strategies	4	2	1	4	4
	Student initiated sensemaking	2	0	3	8	10
	Student initiated check correct	4	2	4	4	10
	Misc. (Uncategorized) Interaction Count	52	85	40	51	58
	Normalized Gain	0.239	0.271	0.190	0.246	0.259

Keeping in mind that the sample size is very small and so generalizations cannot be made from these data, correlations between the types of episodes can be computed to gain a sense of whether the different categories are capturing different types of activities, or whether they are all manifestations of the same type of teaching activity. If IE were a

single activity, we would expect to see high correlations among these categories. This would indicate that an instructor who engages in one type of IE activity is likely to engage in another, and that instructors could be classified as IE or not IE without concern for the particular actions which are taking place. A factor analysis would quantitatively assess how many distinct categories may be represented by these categories by analyzing their correlations. Performing this analysis with a larger sample size may provide further insights into how these types of interactions might be clustered.

We first consider the correlation between instructor-initiated and student-initiated episodes. This correlation is not significant, $r(3) = -0.102$, $p = 0.87$, suggesting that instructors who initiate interactions in their classroom may or may not be more likely to have students ask questions during class. A significant negative correlation might have been observed if each classroom had the same number of total interactions, for example, implying that every interaction initiated by the instructor is one which was not initiated by a student. The actual result suggests, instead, that the number of instructor-initiated episodes and the number of student-initiated episodes might each, separately, classify a classroom, and by extension, the style of instruction in that classroom. This distinction might also be thought of as a classification of the classroom culture, as the four possibilities of no interactions, only instructor-initiated questions, only student-initiated questions, and questions initiated by both the instructor and students describe different classroom environment.

We next consider correlations between the counts of different public interaction categories, displayed in Table 3.

Table 3: Correlations Between Counts of Public Interactions Categories

	Promotes Checks	Encourages revisions from students	Promotes sense making	Problem presented which students have worked on
Encourages revisions from students	-0.114			
Promotes sense making	0.291	0.345		
Problem presented which students have worked on	-0.032	-0.371	-0.474	
Feedback on questions answered by students	-0.216	-0.227	0.464	-0.685

The results in this table are a mixture of positive and negative correlations. This means that while some of the categories, such as promoting sensemaking and feedback on questions answered by students are positively correlated, others, such as problems presented which students have worked on and feedback on questions answered by students are negatively correlated. This indicates that, at least in public instructor-initiated interactions, instructors who are interacting with students are doing so in particular ways, using some types of interactions but not others. While the sample size in this study is too small to make and specific conclusions about trends in instructional style, this type of study along with a factor analysis may allow conclusions to be drawn about specific instructional patterns, determining whether correlated counts of interaction types reflect a smaller number of types of instructional strategies.

If we next consider private episodes, we find that the number of private episodes initiated by the instructor and the number of private episodes initiated by students are not

significantly correlated, $r(3) = 0.065$, $p = 0.92$. This, like the number of public interactions considered by instructor, suggests that the initiators of private interactions may classify types of interactive classrooms instead of simply measuring the degree of interactivity in the classrooms.

We next consider whether there exist correlations between type of public and types of private episodes. These correlations are presented in Table 4.

Table 4: Correlations Between Counts of Public versus Private Episodes

		Private Interaction Types			
		Instructor initiated private episode	Student initiated private episode		
Public Interaction Types		Promotes Checks	0.644	0.254	
		Encourages revisions from students	0.553	0.288	
	Instructor Initiated		Promotes sense making	0.666	0.224
			Problem presented which students have worked on	-0.645	0.636
		Feedback on questions answered by students	0.104	-0.678	
		Student initiated developing strategies	-0.149	0.454	
	Student Initiated		Student initiated sensemaking	-0.794	0.088
			Student initiated check correct	-0.587	-0.436

The results of this table suggest, again, that grouping instructors as IE or TL may ignore some of the subtle differences between instructional styles. More specifically, we notice that the different types of student-initiated public interactions are largely negatively

correlated with instructor-initiated private episodes. It may be that in classes where students actively ask questions during lecture, the students take greater responsibility for their own learning and feel comfortable enough to bring up questions immediately. If this is the case, when engaging in groupwork, the instructor may be less inclined to intervene in their work. If this interpretation is correct, it may be for a variety of reasons. Student confidence may lead instructors to let students work on their own, or instructors allowing students to struggle on their own may lead to greater levels of student self-confidence.

It is also interesting to note that categories of the number of instructor-initiated public interactions are largely positively correlated with instructor-initiated private episodes, but less so with student-initiated private episodes. This may be an indication that instructors who dominate conversation in whole group discussions are also likely to intervene in student groupwork.

The mix of positive and negative correlations indicates that instructors are not falling into a single category of IE or TL, though a future factor analysis with more instructors may help determine how many instructional styles are being expressed in these counts. Instructors who are likely to engage in some types of activities which we classified as IE activities are not engaging in other activities classified as IE. If we take “Student initiated developing strategies” as an example, instructors whose students are likely to frequently propose strategy development are somewhat likely to provide students with more groupwork time, $r(3) = 0.324$, $p = 0.595$, but are unlikely to make comments which promote checks and extend the conversation outside of the immediate problem, $r(3) = -0.551$, $p = 0.335$, though neither relationship is statistically significant. When an

instructor provides students time for groupwork, the available time for other times of interactions decreases. These correlations suggest that those instructors who are not engaging in one type of IE activity, however, are engaging with others. The mix of positive and negative correlations present indicates that considering IE as a single category does not fully describe the instructional practices present in our data, as instructors each use some strategies and not others. This fact is not recognized by simply dividing classes into IE and TL classrooms. This can be seen directly in the results seen in Table 2. Instructor A and instructor E have similar counts for the total number of student-initiated and instructor-initiated episodes (52 and 54 respectively), but these instructors taught using very different styles. Instructor A dedicated substantial time to groupwork, while instructor E used no groupwork at all. Instructor E engaged in far more student-initiated interactions in whole group discussions, however, when compared with instructor A.

3.2 Student Scores on CCI

All the student subjects in the study were required to take the Calculus Concept Inventory as a pretest and posttest as part of their Calculus I semester course, and consenting students had their scores collected along with demographic information such as SAT mathematics scores, university mathematics placement exam scores, gender, and ethnicity. The pretest was graded for course credit on completion, and the posttest scores were factored into a small portion of students' final grades.

There were 26 sections of the course, with a maximum capacity of 35 in each section. Most classes were near capacity, and on average 18.5 students per section participated in

the study, ranging from 10 to 26.

The classrooms of the 5 instructors who agreed to participate represented a spectrum of normalized gain scores on the CCI ranging from 0.19 to 0.27. The mean normalized gain for the entire participant group at the large, southwestern university where our study was conducted was 0.25, meaning that 25% of the previously unknown concepts was learned during the course. Normalized gain scores for the entire 26 sections ranged from 0.14 to 0.36. Most of these classes, including some of those within the instructor participant sample, had normalized gains scores of about 0.20, which is near the national average.

In his 1998a study, Hake categorized introductory mechanics classes by their normalized gain, $\langle g \rangle$, scores: “low- g ” sections were defined as those with $\langle g \rangle$ values less than 0.3, “medium- g ” as those between 0.3 and 0.7, and “high- g ” as those above 0.7 (p. 65). By Hake's definition, 4 of the 26 sections had medium- g scores, and the remaining 22 had low- g scores. By comparison, the scores from the University of Michigan reported by Rhea (n.d.) ranged from the low to medium- g scores, with an average at the low end of the medium- g range. The highest scores reported there were between 0.40 and 0.44, which is still substantially less than the necessary $\langle g \rangle$ score of 0.7 for a high- g classroom.

3.3 Normalized Gain Scores on the CCI

In his study, Epstein (2007) gave the CCI to 1100 students at 12 American universities and 1 university in Finland. He found $\langle g \rangle$ values largely clustered between 0.15 and 0.23, similar to the scores in traditional lecture physics classes on the FCI. A

large mid-western research university with a department-wide IE-focused teaching style reported an average $\langle g \rangle$ score of 0.35 among their 51 sections, with a range of 0.21 to 0.44 (Rhea, n.d.). Ten of the sections had $\langle g \rangle$ scores above 0.40.

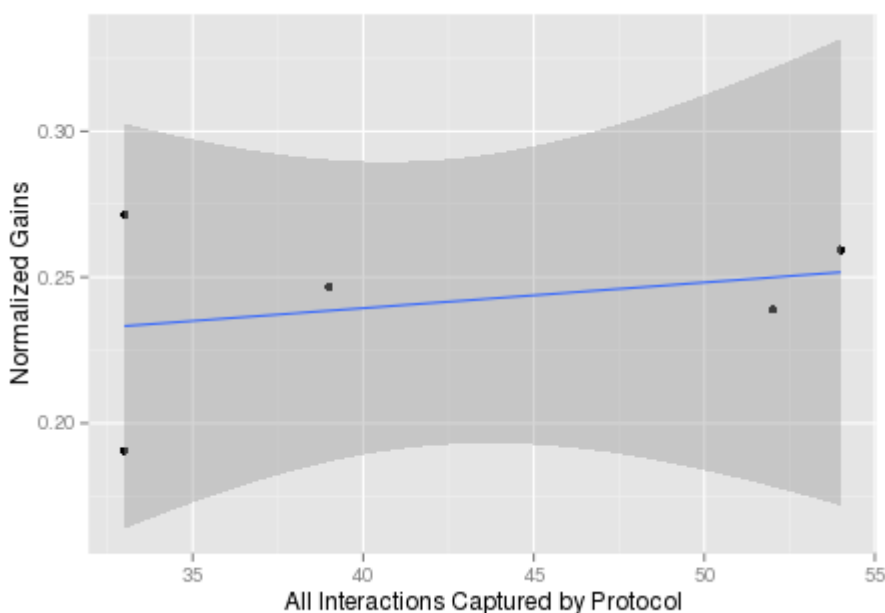
3.4 Results of Comparing the Coding Protocol with CCI Scores

3.4.1 Main Hypotheses

We are interested in knowing whether the number of interactions predicts student gains. We consider the total number of interactions captured by the coding protocol initially, and then consider only the student-initiated episodes or instructor-initiated episodes separately.

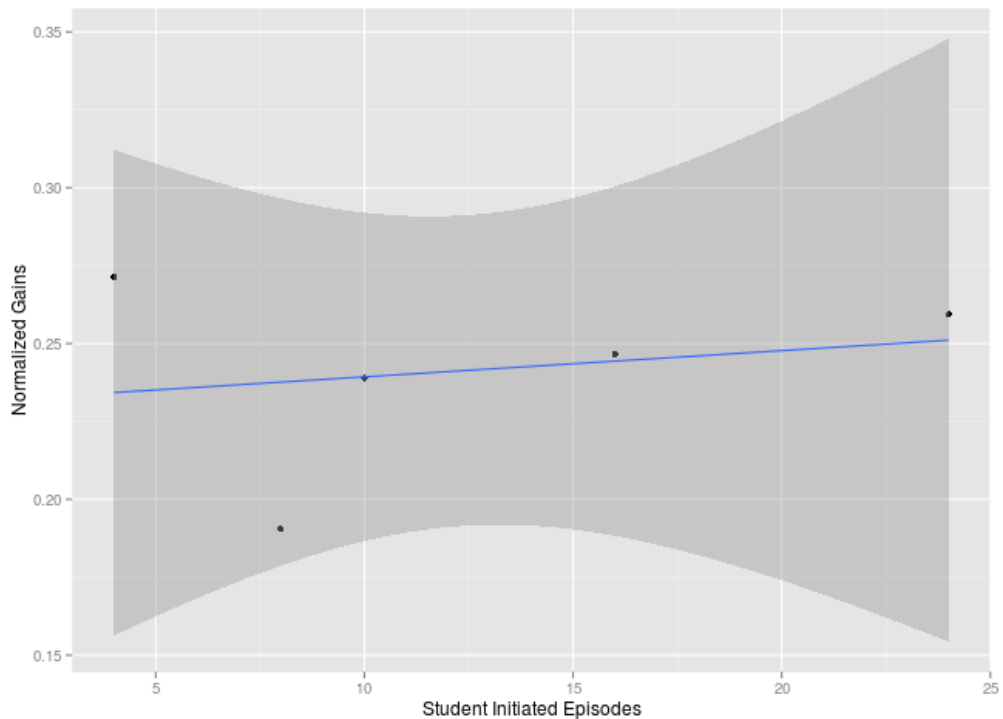
Students' normalized gain scores on the CCI are not predicted significantly by the total number of interactions captured by the categories defined in the protocol. Results are displayed in Table 5 as model 1, and are graphed in Figure 1. We note that this does not include the interactions categorized as “miscellaneous” in the protocol.

Figure 1: Normalized Gains versus Protocol Interactions



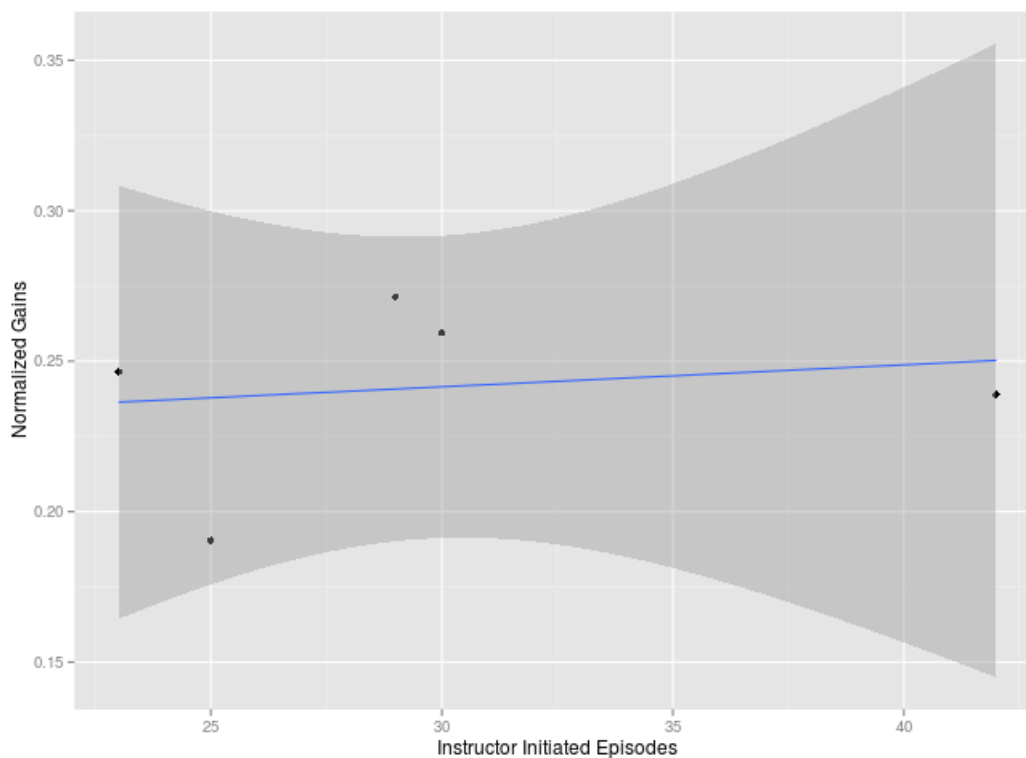
We then use a linear regression to find that the number of student initiated episodes does not significantly predict student gains, with results displayed in Table 5 as model 2, and plotted in Figure 2.

Figure 2: Normalized gains versus student initiated episodes



Care needs to be taken in interpreting the results of this regression, however, as the two classes with low counts of student initiated episodes do not follow the pattern seen in the other three classes. Since a potential outlier exists in such a small sample, any trends observed should be considered tentative until a larger sample indicates whether this class is an outlier or evidence of a more complicated pattern. This is then compared with a regression using instructor-initiated episodes to predict student gains, also not significant. Results are displayed in Table 5 as Model 3 and are plotted in Figure 3.

Figure 3: Normalized gains versus instructor initiated episodes



In this plot, the relationship between instructor-initiated episodes and classroom normalized gains is even less pronounced. It is not possible to support the claim that the total number of student-initiated or instructor-initiated episodes is correlated with student gains based on these data. Given the small number of instructors observed, however, it is also not possible to say whether this is due to a lack of statistical power or the lack of a genuine relationship. Replication with a larger number of instructors could help to differentiate between the two possibilities.

Table 5: Regressions of Normalized Gains by Counts of Interactions

Variables	B	SE(B)	β	t (df = 3)	Sig (p)	R ²
Model 1: Total Number of Interactions Captured by Protocol						
Constant	0.204	0.073		2.810	0.067	0.083
All Interactions	0.001	0.002	0.288	0.521	0.638	
Model 2: Student-Initiated Episodes						
Constant	0.231	0.032		7.229	0.005	0.045
Interactions	0.001	0.002	0.211	0.375	0.733	
Model 3: Instructor-Initiated Episodes						
Constant	0.220	0.073		3.016	0.057	0.030
Interactions	0.001	0.002	0.174	0.306	0.780	

Note: B indicates the unstandardized regression coefficient. β indicates the standardized regression coefficient.

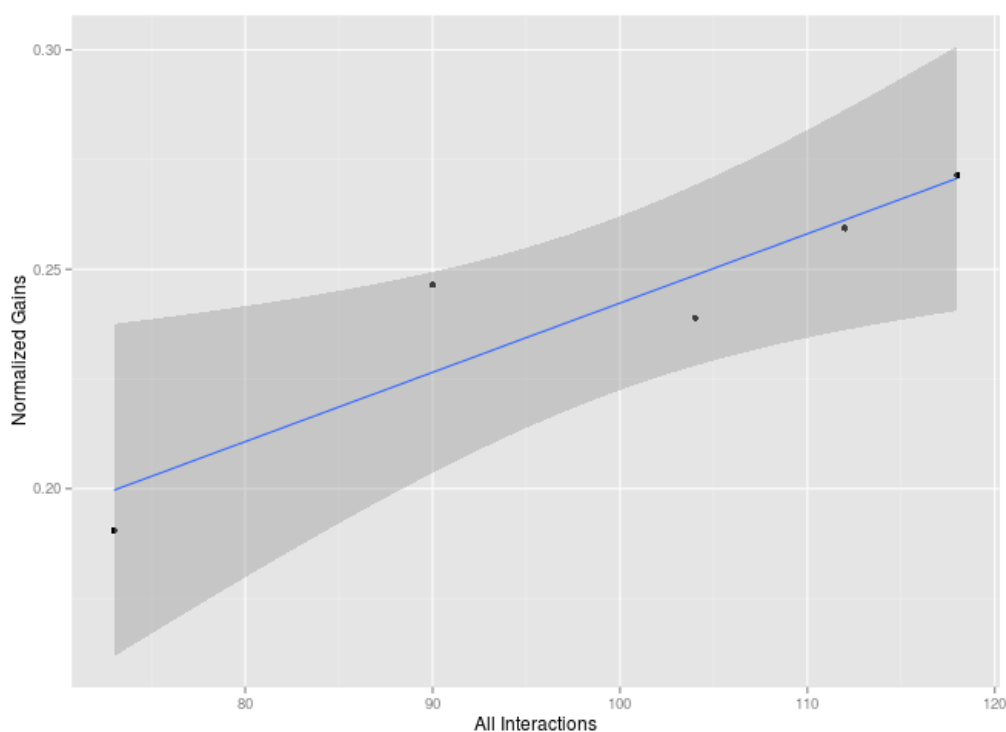
In Table 5 we see that none of the types of interactions significantly predict normalized gain scores as all p -values for interactions are well above 0.05. We also consider the effect size for each of these models by computing Cohen's f^2 statistic. All three of the models are seen to have small effect sizes of 0.09, 0.05, and 0.03 respectively, meaning that very little variance is explained by these models.

3.4.2 Exploratory Analysis

Since the correlations between the total number of student-initiated or instructor-initiated episodes and CCI instructor-level gains are not significant, we explore whether other methods of measuring IE based on the protocol correlates with student gains. For example, the category of “Miscellaneous Interaction” counted the number of interactions

which did not fit into the predefined categories. By combining this category with the counts in the predefined categories, a total number of interactions, called “All interactions” is created, and indicates the level of IE in the classroom in a different way than the coding protocol does as these were not included in the protocol. This may be evidence of some other variable such as “classroom atmosphere.” This relationship between this count and gains on the CCI is significant, with results displayed in Table 6 as Model 1, and plotted in Figure 4.

Figure 4: Normalized gains versus all interactions



If we explore specific types of interactions, encouraging revisions has a strong correlation with student gains on the CCI. This relationship is also significant, with results displayed in Table 6 as Model 2 and plotted in Figure 5.

Figure 5: Normalized gains versus number of revision encouragements

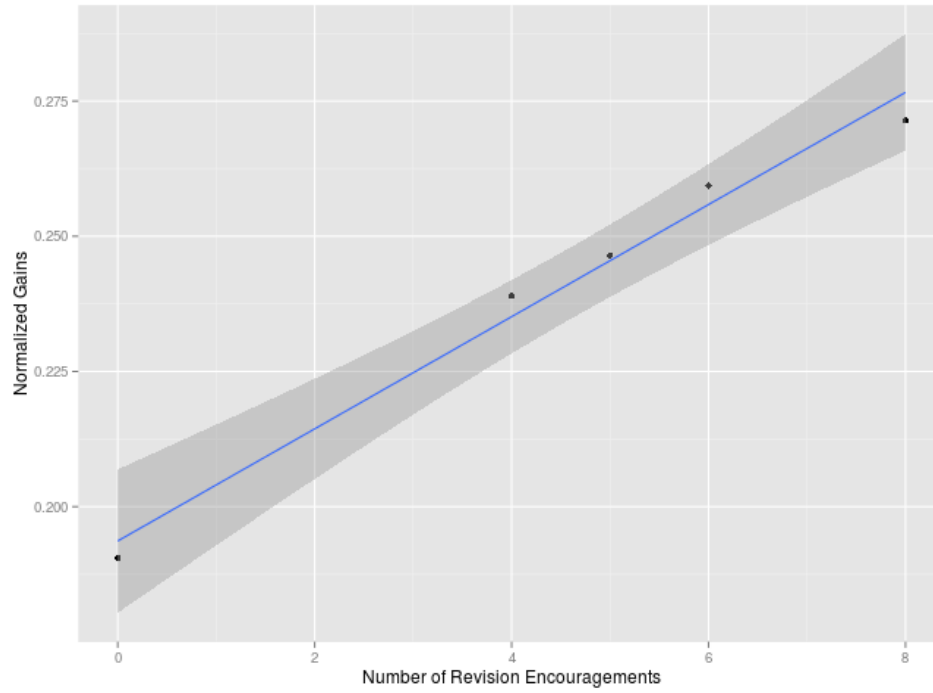


Table 6: Regression Results for Exploratory Analysis

Variable	B	SE(B)	β	t (df = 3)	Sig(p)	R ²
Model 1: All Interactions						
Constant	0.084	0.039		2.181	0.117	0.849
Interactions	0.002	0.0003	0.922	4.108	0.026	
Model 2: Number of Revisions Encouraged						
Constant	0.194	0.004		46.49	< 0.001	0.983
Interactions	0.010	0.001	0.992	13.22	< 0.001	

Note: B indicates the unstandardized regression coefficient. β indicates the standardized regression coefficient.

The effect sizes as measured by Cohen's f^2 statistic for the two models are 5.63 and 58.17

respectively, which are very large. While we might expect such a large R^2 in cases where we are measuring the same quantity in different ways, this is not the case with numbers of interactions and gain scores. Given the small sample size, we determined that while these results do suggest that a relationship between these two types of interactions and student gains may exist, the magnitude of the R^2 values and effect sizes should be taken as an artifact of this sample and not necessarily as an indication of the actual strength of the relationship. Additionally, the computation of R^2 presumes that there is no measurement error. In our study, while over 80% reliability was achieved in videocoding, this source of error should be kept in mind when considering these values of R^2 and f^2 , so there are additional sources of error not considered in these statistics.

4 Conclusions

The analysis of this chapter suggests that the coding protocol differentiates styles of instruction, since the correlations between different types of interactions were sometimes positive and sometimes negative, and both significant and non-significant. It seems unlikely based on this evidence that IE instruction is a single activity which instructors engage in, and instead consists of many types of activities, meaning that this coding protocol contributes something not currently in the research literature and provides additional characterizations of IE instruction beyond a dichotomous variable. Instructors may engage in all of these activities, none of them, or some of them. Given a larger sample of instructors, it may be possible to conduct a factor analysis to determine whether certain types of interactions are evidence of particular teaching practices. This may allow us to classify types of IE instruction and determine whether classifications are

associated with higher levels of student gains. While the interaction types seem to differentiate instructional styles, the coding protocol categories individually do not significantly predict CCI gains, except for the number of revisions encouraged. Additionally, the total number of interactions which occurs during a class significantly predicts CCI gains. If this total number were a better measure of interactivity than just the IE categories captured by the protocol, then overall interactivity may be a gateway for students to become engaged with the material. It is also possible that IE episodes which encourage calculus knowledge should not be restricted to those IE episodes which focus on calculus content. Further expansion of the IE categories to precalculus material seems to be a reasonable consideration since the development of precalculus ideas could support the underlying knowledge of topics necessary to understand the ideas of calculus. For example, if students have poor understanding of the concept of rate of change, a conceptual understanding of derivatives would be difficult to achieve. This result is similar to results found by Deshler (2009), who found that students in classrooms labeled as “highly interactive” were more likely to be successful in the class. “Highly interactive” classrooms were largely determined by the number of student-teacher interactions. This study shows that the results Deshler found seem to extend to tests of conceptual understanding in addition to traditional measures of student success.

Furthermore, the specific category of “Number of revision encouragements” is also significantly correlated with student gains, suggesting that at least this one specific type of interaction may be particularly helpful for fostering students' conceptual gains. We do again suggest that interpretation of the effect size and R^2 values be tempered by the

understanding that the videocoding process introduced measurement error which is not taken into account here.

CHAPTER 3: HIERARCHICAL LINEAR MODELING AND ANALYSIS OF INDIVIDUAL-LEVEL PREDICTORS

In this chapter, we introduce a statistical technique called hierarchical linear modeling (HLM), also known as multi-level modeling. HLM is useful for analyzing data which is clustered in groups, such as students within classrooms or patients within hospitals. Systematic differences between classrooms or hospitals can result in differences between groups which are the result of clustering. For example, consider a clinical trial being administered in multiple hospitals. The different hospitals may administer a drug in different ways, for example, leading to differences between the hospitals which are not related to the effect of treatment. In a hierarchical linear model for this example, differences between hospitals are considered “random effects” and the remaining “fixed effects” can be attributed to the effect of the clinical trial. Hierarchical linear models can include variables at multiple levels, so our example might include predictor variables at the individual level, such as gender, as well as at the hospital level, such as environmental variables for the area, which would be the same for all individuals at that particular location.

Because we have students nested within classrooms, we have two levels in our hierarchical linear model (Raudenbush, 1993). At level one, we are predicting student level outcomes. To model the score for student i in classroom j , we use the model given by Equation 2.

$$y_{ij} = \beta_{0j} + \sum_{q=1}^Q \beta_{qj} x_{qij} + r_{ij}, r_{ij} \sim N(0, \sigma^2) \quad (2)$$

In this equation, x_{qij} are the student level predictors, where j and i designate the classroom and student respectively, and q ranges over the Q student-level predictors. β_{0j} is the intercept for classroom j , and β_{qj} are the regression coefficients for classroom j 's Q predictors. Note that each classroom has its own set of regression coefficients and its own intercept. Finally, r_{ij} represents the error terms in the model.

The hierarchical linear model also includes a level two component, to model at the classroom level. At this level, we model the β_{qj} terms used above in Equation 2, by using Equation 3.

$$\beta_{qj} = \Theta_{q0} + \sum_{s=1}^S \Theta_{qs} W_{sj} + u_{qj} \quad (3)$$

In this equation, q again designates the Q student level predictors, and j represents the classroom. Therefore, for a fixed value of j , this equation combines the intercept Θ_{q0} , the classroom level characteristics W_{sj} , their regression coefficients Θ_{qs} , and error terms u_{qj} to predict the regression coefficients for the level one equation, Equation 2. A complete treatment of HLMS can be found in Gelman and Hill (2007) or Raudenbush and Bryk (2002).

In this chapter, I will use the terms “individual-level” and “instructor-level” to refer to the variables which will vary between individual students and those which are the same for all students in the same class respectively. In this study, the instructor-level variables are the variables determined through the coding protocol, and the grouping is done by the

instructor. By using an HLM, we are also able to use individual pretest and posttest scores to create gain scores at the individual level instead of at the instructor-level, as was done in the previous chapter.

Using an HLM has two advantages over the methods used in the previous chapter. First, as mentioned earlier, creating an HLM allows us to use individual-level covariates in the model. This allows us to build a model which takes more information into account and allows for investigation into whether certain instructional practices are more beneficial for some students than others. For example, by taking gender into account, we can determine whether IE instructional techniques encourage gains among female students more than male students. Second, when the clustering is taken into account, we can use an HLM to determine the proportion of variance in the scores which can be explained at the individual-level and at the instructor-level. The analysis in the previous chapter only considered the average gains for each instructor. We are only able to compare class averages with each other, but have no information about the distribution of scores within a single instructor.

Research suggests that individual factors such as gender are related to success in STEM fields (Seymour & Hewitt, 1997). Specifically, research in physics education has shown that IE teaching techniques can reduce and, in some cases, eliminate gender gaps (Kost et al., 2009; Lorenzo et al., 2006; Miyake et al., 2010; Pollock & Finkelstein, 2007). In calculus, classes designated as Inquiry-Based have shown results in reducing gender gaps using traditional measures of learning and in affective variables (Laursen et al., 2011). Hake (1998) also suggests that individual-level variables may explain

additional differences in student scores. These results suggest that individual-level variables such as gender are important to consider, and so instructor-level analysis is insufficient for our analysis.

1 Individual Normalized Gains

Most analyses of concept inventory results use class averages to calculate normalized gain scores. This choice, when used with a measure of IE, implies an assumption that these public interactions will have class specific effects. Since instructors are not likely to treat all students equally, this assumption may not be valid, and it may be that students who are engaging in the interactions are benefiting from the interactions while other students are not. In this case, it would be beneficial to note which students were engaged in the interactions and determine whether their gains were higher than other students in the class. Our protocol did not distinguish between five interactions with the same student and five interactions with five different students. Despite this, there is also reason to think that IE instruction may have class-wide benefits. It may be the case that the type of student thinking encouraged in an IE classrooms is beneficial to all students in the classroom. For example, for an instructor who frequently encourages revisions, asking the class to question or revise a statement may encourage all students to consider and engage with the question internally, regardless of which student or students respond to the question. If this is the case, it is actually the question-asking which is of the highest importance, and the effect would be class-wide. It may also be the case that IE instruction is evidence of a classroom which encourages student learning instead of the cause itself. For example, a classroom in which students are frequently asking questions may be a

classroom in which students feel comfortable enough with the instructor and the content to ask these questions. In this case, asking the questions is only a symptom of the classroom environment which is beneficial to the students. If this is the case, this classroom environment may be beneficial to more students that are those interacting, as it is the classroom atmosphere which is beneficial, if not to all the students in the classroom.

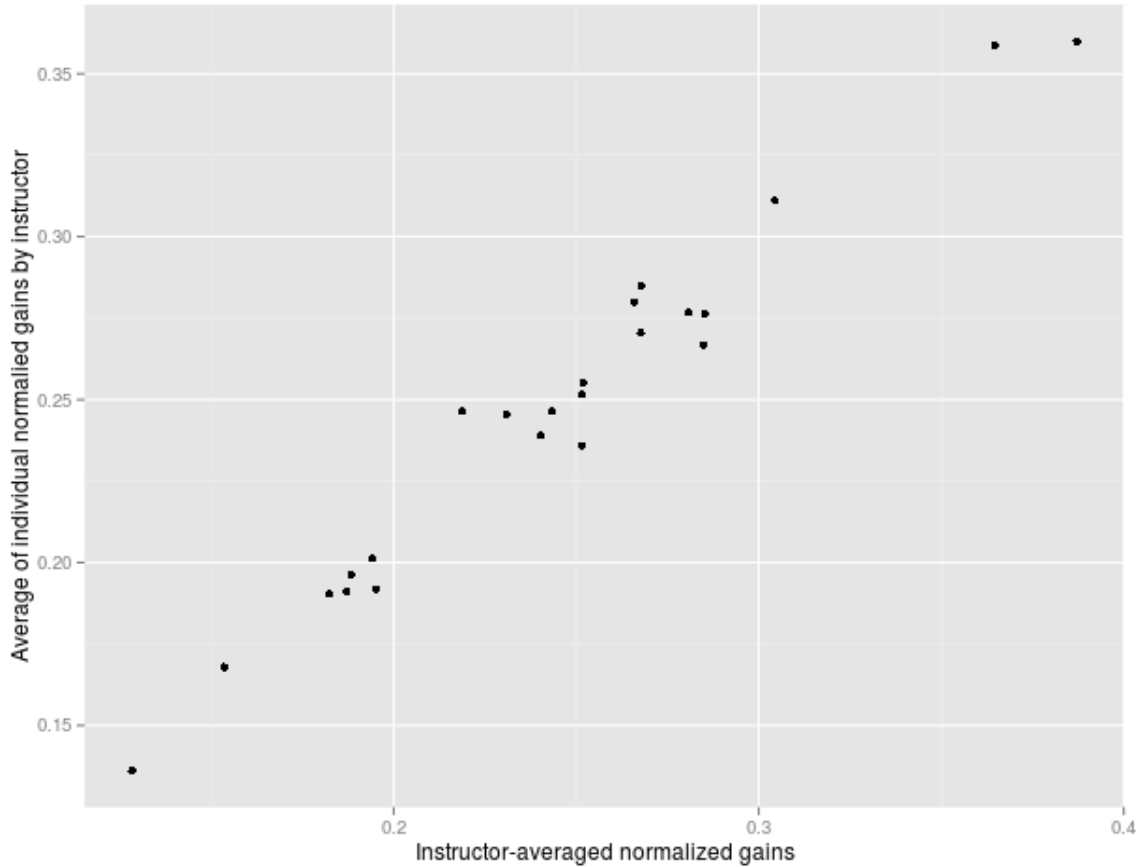
By computing these gains in the typically used way, we were able to determine how much the mean score of a class changes when IE techniques are used. One can also compute normalized gain scores at the student level by using individual pretest and posttest scores. Averaging these individual-level normalized gains can sometimes produce results which are different from normalized gains calculated using class-average pretest and posttest scores (Bao, 2006; Coletta & Phillips, 2005). Bao (2006, p. 919) explores theoretical examples, in which different $\langle g \rangle$ scores can be obtained depending on whether gains are uniform among all students resulting in a uniform shift of all students, or whether gains are different for different students resulting from a change in the rank-ordering of the students. Inequalities can be found for particular scenarios, but Bao finds that neither method results in scores which are always greater than the other (p. 921).

If we introduce individual level predictors such as gender or SAT mathematics scores in addition to the instructor-level IE measures, we can build an HLM predicting individual-level normalized gain scores based on both individual-level characteristics and instructor-level IE variables. These normalized gain scores are calculated using the same

formula for $\langle g \rangle$ given in Chapter 2, only using individual pretest and posttest scores instead of class averages.

In our data, the difference between using individual normalized gains averaged by class and class average normalized gains makes very little difference. The mean of the average of individual level normalized gains ($M = 0.2450$, $SD = 0.061$) is slightly lower than the mean of the class-average normalized gain scores ($M = 0.247$, $SD = 0.055$), though very close. Bao (2006) does not make any specific recommendations about when one measure should be used. He instead uses simulated data to show that under some specific conditions, it is possible to achieve different results depending on which method of calculating normalized gains is used. A scatterplot of normalized gain scores computed by averaging individual-level normalized gains versus class-average (traditionally computed) normalized gain scores is given in Figure 6. The measures of gain are highly correlated, $r(21) = 0.983$, $p < 0.001$, and it appears that, in this case, there is little difference depending on which measure is used.

Figure 6: Comparison of classroom level and average individual normalized gains



2 Null Model

The typical first step when building an HLM is to construct a model called a null model, which estimates the variance at the instructor-level and at the individual-level without any predictor variables (Roberts & Monaco, 2006). Constructing null models allows one to determine what percentage of variance lies in each of these levels, which allows one to understand how well variables at each level might explain the gain scores. For example, if the percentage of variance at the instructor level is low, then IE variables are unlikely to have much explanatory power. In a normal regression, an analogous model would have no predictors, so the regression would only predict the mean of the

outcome variables.

For our null model, we used all students who took the CCI ($N = 482$) instead of only those who were taught by instructors who were videotaped ($N = 130$). Since we are interested in how the variance splits, we used the largest possible data set that we can. For the null model, we used every data point that has a normalized gain score, which is all students who took the CCI twice. Once the division of variance in the total population has been determined, we can determine whether the set of students who were taught by instructors who were videotaped is similar, and use both individual and instructor-level covariates.

The results of the HLM are seen below in Table 7.

Table 7: Results of HLM Null Model

<i>Fixed Effects</i>	Null Model
Intercept	0.246
(SE)	0.010
<i>t</i> -value	23.5
<i>Random Effects</i>	
Intercept Variance	6.1847e-13
Residual Variance	5.2710e-02

The fraction of variance attributable to the instructor level is computed by finding the ratio of variance at the instructor level ($6.1847e-13$) to the total variance ($6.1847e-13 + 5.2710e-02$). This means that nearly all the variation (over 99.9%) exists at the individual level. These results suggest that any instructor-level variables, namely degree of IE, are able to explain less than 0.1% of the variation in the data, and that individual variables such as gender could explain the remaining variation. The variation between students of different instructors is very small compared with the variation which exists among all the

students, suggesting that there is very little difference between classrooms. If IE techniques were truly improving gain scores among students, we would expect instructors with higher IE levels to have students with gain scores much higher than those with low IE levels, even when the variance at the student level is taken into account. If this were the case, the variance in student gains would be largely explainable by the student's instructor and not by any residual effects which lie at the individual student level. It is also possible that prior studies have incorrectly interpreted a spurious relationship. The discrepancy between results when classroom scores are aggregated and when HLM is used illustrates the need for care when analyzing concept inventory gains. We found that the relationship which seemed to exist between the level of IE instruction and student gains appears to be spurious, but only when variation in score at the student level is included in the model.

While we find no evidence for a relationship between IE instruction and student gains on the CCI at the individual student level, we cannot claim that the relationship does not exist. It is possible that the relationship is real but we are not able to see evidence for it. For example, it could be the case that the effect of interactive teaching is only seen after a certain level of interactivity occurs. If this is the case, it may be that the instructors at this university were not teaching at the level of interactive engagement which would encourage the types of gains seen by Rhea (n.d.). From this study, we cannot determine whether frequency of IE episodes, types of IE episodes, or a combination is important for encouraging student gains, as the interpretation of the results of the null models suggests that no instructor-level variables will be able to explain much variation in scores. This

may be a productive area for future research, where collecting data from a set of instructors with more diverse teaching styles may help to answer this question. It is also possible that the variety of the instructional styles in our data was not diverse enough to determine the effect of instructional practices. If a greater variety of instructional practices were observed, we may have seen more drastic differences between classrooms and could have attributed instructor-level variables to these differences. These two possibilities are similar, though describe different possible instructional realities. The first possibility suggests that no effects are seen until a “cutoff” level of interactivity, so that plotting student gains against instructional interactivity levels would have the appearance of a step-function. The second suggests that a linear regression may in fact describe the relationship between interactivity levels and student gains, but the variety of instructional styles may need to be more diverse than is seen in a single university. A larger study including multiple universities may provide us the ability to determine whether instructional style, perhaps at the university level, affects student gains on the CCI.

A hierarchical linear model predicting individual CCI gains is likely to produce results nearly identical to that of a linear model which ignores instructor-level grouping. Since there is so little variance explainable at the instructor level, the instructor-level grouping provides very little information. As nearly all the variation lies at the individual level, the variables which are likely to explain differences are all at the individual-level, suggesting that the clustering that HLM uses will provide no additional benefits.

It is unclear whether previous studies which only aggregate data at the instructor-level would have similar results if the within-classroom variance were considered in the

models. Since reports by those such as Hake (1998b), Rhea (n.d.), and Epstein (2007) along with many others only report class averaged gains, it is possible that some of the differences observed between classrooms with different instructional styles may be less pronounced. Additionally, by considering individual level gains, studies by the authors mentioned in addition to many others may provide insights as to whether certain instructional techniques are more effective at encouraging gains among certain types of students by considering individual-level variables.

3 Pretest and Posttest Analysis

Normalized gain scores were considered as the outcome variable in the previous model. We also consider whether significant differences exist in pretest and in posttest scores. This helps us to better understand whether students, who are all achieving roughly the same gains regardless of instructor are also starting with and ending up with roughly the same level of skill. Students in all classrooms are achieving roughly the same gains when compared with the variety of individual student gains, but this is based on gains measured in normalized gains. It may be that students in different classrooms have very different pretest and posttest scores from each other, but the specific construction of the normalized gain score equates them. By predicting posttest scores, we have the ability to determine how individual-level variables or instructor-level variables affect posttest scores while controlling for pretest scores, provided variation exists in pretest and posttest scores between instructors.

We create two null models, one using pretest scores as an outcome variable and one using posttest scores as an outcome variable. While the pretest scores are not being

predicted in any meaningful way, the two null models allow us to determine what fraction of variance lies at the individual-level and what fraction lies at the instructor-level. Analysis of pretest scores using a null model indicates that approximately 5% of the variance lies at the instructor level, and analysis of posttest scores indicates that approximately 5% of this variance is at the instructor level. An analysis of variance (ANOVA) illustrates that while no statistically significant differences in normalized gains between instructors is seen, $F(22,459) = 1.39, p = 0.18$, differences are seen between instructors on the pretest, $F(22,459) = 2.105, p = 0.003$, and between instructors on the posttest, $F(22,459) = 2.102, p = 0.003$.

This suggests that predicting posttest scores with an HLM, using pretest scores as a predictor variable, may allow us to determine whether IE teaching techniques are encouraging gains. When we restrict our data set to students of those instructors we have observed, however, the differences between instructors is not apparent. An ANOVA shows that there is no longer a statistically significant difference in pretest scores among those instructors, $F(4,125) = 1.236, p = 0.299$, nor is there a statistically significant difference in posttest scores, $F(4,125) = 0.73, p = 0.573$. Running null models suggests the same result, as a null model predicting pretest scores with clustering by instructor indicates over 99.9% of the variance is at the individual-level, and over 99.9% of the variance in posttest scores is also at the individual-level.

These results indicate that, while small, there may be differences in posttest scores among all the students which could be attributable to instructor-level differences, though the differences between scores do not exist among the instructors observed as part of this

study.

4 Analysis of Gender

As already mentioned, results in physics have suggested that IE instruction may help to reduce gender gaps more effectively than traditional classes (Kost et al., 2009; Pollock & Finkelstein, 2007) and results in Inquiry-Based calculus classes suggest similar results (Laursen et al., 2011). For this reason, gender was considered a variable of interest, and we construct a model predicting gains using gender as a covariate. Whenever possible, the full data set ($N = 482$) is used. This is the case whenever instructor-level (IE) variables are not included in the model. When instructor-level (IE) variables are included in the model, the smaller sample size of students whose instructors were observed ($N = 130$) is used.

On the CCI pretest, male students outperformed female students by 1.13 out of 22 questions on average, $t(470.148) = 3.94$, $p < 0.001$, and this relationship was also seen on the posttest, with male students outperforming female students by 1.33 questions, $t(421.426) = 3.79$, $p < 0.001$. While these results suggest that there is a gender gap both at the beginning and the end of the semester, there is no significant difference in normalized gains between genders, $t(426.062) = 1.42$, $p = 0.157$. This suggests that while a gender gap does exist, it is not changing significantly over the course of the semester. This is important because results in physics (Kost et al., 2009) and mathematics (Laursen et al., 2011) have shown that interactive teaching techniques can diminish gender gaps. If this is the case, we would expect that the gender gap may be decreased in the classrooms of instructors who implement IE instructional techniques. The results observed here may be

a result of a lack of full IE implementation or a lack of enough variation in the instructional styles among the instructors observed. If this result is for either of these two reasons, this may be evidence that increasing the level of IE in a class only a little does not correlate with gains, and perhaps a critical amount of interactions are needed before student gains begin to be observed.

It is possible that an HLM which contains instructor-level variables, student-level variables, and interactions between the two types could contain significant results, even considering the lack of variance at the classroom level. For example, consider a set of classrooms in which instructors implemented IE techniques not at all or very fully. Suppose that in the high IE classrooms, high gains were achieved by female students and low gains were achieved by male students, while in the low IE classrooms the reverse was true. On average, there would be no difference in gains between the classrooms because the gain scores between classrooms would look identical when gender is ignored. Further, if the classroom groupings and instructor-level variables (IE) are ignored, female students and male students would be observed to be achieving roughly the same gains as female students are achieving both high and low gains depending on which type of classroom they are present in, and the same is true for male students.

We built a model using all students who were in classes taught by observed instructors ($N = 130$) which includes both gender and a measure of IE. Because this model includes both instructor-level variables (level of Interactive-Engagement) and individual levels variables (gender), an HLM is necessary. Since the number of revisions encouraged was a statistically significant predictor of student gains when scores were

aggregated in chapter 2, we use this as our first measure of level of IE. The results of the model are given in the following table. The outcome variable in this model is student-level normalized gains ($N = 130$), and the results are presented in Table 8 as Model 1. We expect that gender will likely not be a significant predictor of gains given the results of the t -test described earlier in this section. We also expect that the IE variable (Number of revisions encouraged) is not likely to be a significant predictor of gains given the small amount of variance at the instructor-level. We are interested in determining whether the interaction between the two variables is significant. If it is, it may suggest that IE instructional techniques are more beneficial for one gender than the other.

The only significant variable is the intercept, suggesting that none of the predictors significantly predict gains. In particular, the lack of significance in the interaction term, Gender by number of revisions encouraged, suggests that IE instruction is not affecting students of different genders differently in our data. To illustrate this point, we consider the modeling equations for each gender, where males are given by Gender = 0 and females are given by Gender = 1 in the above table:

$$\text{normalized gain}_{\text{males}} = 0.248896 + 0.002389 * (\text{number of revisions encouraged})$$

$$\text{normalized gain}_{\text{females}} = 0.125229 + 0.022288 * (\text{number of revisions encouraged})$$

While these terms are not significant, the model suggests that each revision encouraged correlated with a gain in normalized gain score, though this increase is roughly ten times larger for female students than male students. This result, while not statistically significant, suggests that, along with the results of other researchers in STEM (Hazari, Tai, & Sadler, 2007; Kost et al., 2009; Laursen et al., 2011; Miyake et al., 2010; Tartre &

Fennema, 1995), further investigation into the relationship between IE instruction and gender may be fruitful for future studies. By setting the two equations above equal to each other, we can see that the gender gap could be reduced by as few as 6 revision encouragements over 3 classes, or 2 encouragements per hour of class. This claim should be taken as only an indication of possible future results, as the variables are not statically significantly correlated with gains, and so further studies would need to be conducted to determine whether an increased sample size of instructors would lead to statistically significant results.

In addition to the number of revisions encouraged, the total number of interactions was found to be significantly correlated with student gains in Chapter 2, so we construct an HLM predicting individual normalized gain scores using gender, the total number of interactions, and the interaction between gender and all interactions. This model includes all students who were in classes taught by instructors who were observed (N = 130). The results of the model are presented in Table 8 as Model 2.

Table 8: HLMs Predicting Individual Normalized Gains

<i>Fixed Effects</i>	Model 1	Model 2
Intercept	0.249	0.174
(SE)	0.056	0.165
<i>t</i> -value	4.458	1.057
Gender	-0.124	-0.252
(SE)	0.078	0.235
<i>t</i> -value	-1.585	-1.075
Number of revisions encouraged	0.002	
(SE)	0.010	
<i>t</i> -value	0.247	
Gender by Number of revisions encouraged	0.020	
(SE)	0.014	
<i>t</i> -value	1.431	
All interactions		0.001
(SE)		0.002
<i>t</i> -value		0.538
Gender by All interactions		0.002
(SE)		0.002
<i>t</i> -value		0.978
<hr/> <i>Random Effects</i> <hr/>		
Intercept Variance	< 0.001	< 0.001
Residual Variance	0.045	0.045
% Variance Explained	< 0.001	< 0.001

We find similar results to using the number of revisions encouraged as a measure of IE, as none of the variables is a significant predictor of student level gains.

One final consideration is whether the choice of using the normalized gain score is affecting the results presented in the previous two models. Before considering the effects of IE as in the previous two models, we first consider a model predicting posttest scores using gender, pretest score, and the interaction between the two as predictor variables. The interaction term provides us with information about whether incoming knowledge

affects posttest scores in different ways for different genders. For example, if the interaction term is significant and positive, that would imply that the slope of the line predicting posttest scores from pretest scores is steeper for female students than male students. Since there are no instructor-level variables involved, this regression can be run as either a typical ordinary least squares regression or as an HLM. Given the lack of variance explained at the instructor-level, we do not expect these models to product different results. To illustrate this fact, we present the results of the standard regression ($N = 482$) in Table 9 and the results of the HLM in Table 10.

Table 9: Regression Predicting Posttest Scores from Pretest Scores and Gender

Dependent variable: CCI Posttest Score					
Variable	B	SE(B)	β	t	Sig. (p)
(Intercept)	6.151	0.495		12.420	< 0.001
Pretest score	0.693	0.052	0.586	13.240	<0.001
Gender	-1.047	0.795	-0.135	-1.317	0.188
Pretest score by Gender	0.064	0.092	0.054	0.698	0.486

Note: $R^2 = 0.3836$, (Adjusted $R^2 = 0.3797$). B indicated unstandardized regression coefficient. β indicates standardized regression coefficient.

Table 10: HLM Predicting Posttest Scores from Pretest Scores and Gender

<i>Fixed Effects</i>	Null Model	Model 1	Model 2	Model 3
Intercept (SE) <i>t</i> -value	11.724 0.254 46.090	6.151 0.495 12.420	5.487 1.046 5.248	5.097 2.345 2.174
Pretest score (SE) <i>t</i> -value		0.69316 0.05235 13.240	0.759 0.083 9.177	0.752 0.083 9.062
Gender (SE) <i>t</i> -value		-1.047 0.795 -1.317	-1.634 1.071 -1.526	-3.403 3.223 -1.056
Pretest score by Gender (SE) <i>t</i> -value		0.064 0.092 0.698		
Number of revisions encouraged (SE) <i>t</i> -value			0.003 0.133 0.023	
Gender by Number of revisions (SE) <i>t</i> -value			0.257 0.191 1.343	
All Interactions (SE) <i>t</i> -value				0.004 0.022 0.207
Gender by All Interactions (SE) <i>t</i> -value				0.030 0.032 0.948
<i>Random Effects</i>				
Intercept Variance	0.762	< 0.001	< 0.001	< 0.001
Residual Variance	14.051	9.140	8.404	8.489
% Variance Explained		< 0.001	< 0.001	< 0.001

We see that while pretest scores significantly predict posttest scores as expected, neither gender nor the interaction term significantly predict posttest scores. This suggests that the relationship between pretest score and posttest score are independent of gender.

We now consider the possibility that the choice of normalized gain score may have

affected the first two models of this chapter. To address this possibility, we consider each of these models in which we replace the predicted variable of normalized gains with posttest scores and include pretest scores as a predictor variable ($N=130$). The first model predicts posttest scores using pretest scores, gender, and the number of revisions encouraged as a measure of IE, while the second model predicts posttest scores using pretest scores, gender, and the total number of interactions as a measure of IE, and the results are reported in Table 10 as Model 2 and Model 3 respectively. These results show that there are no different interpretations when predicting posttest using pretest and additional variables from predicting normalized gains with those additional variables. This provides us with evidence that the particular method of measuring gains does not seem to be affecting the results presented in this study.

5 Analysis of Previous Mathematics Experiences

While previous studies have considered the relation between gender, IE instruction, and concept inventory gains in physics, for example by Kost et al. (2009), other individual-level variables might affect CCI gains, or how effectively IE instruction is at encouraging student gains on the CCI. The individual-level variable which we investigate in this section describes prior mathematics classes students have taken, and whether these classes were taken in college or in high school. To achieve this, students marked responses to the following questions:

1) Have you taken calculus previously?

No _____ Yes, in high school _____ Yes, in college or university _____

2) Have you taken pre-calculus previously (functions, trigonometry, advanced algebra)?

No _____ Yes, in high school _____ Yes, in college or university _____

We initially began to investigate these data ($N = 482$) by considering each possible blank to indicate a dichotomous variable. For example, “no calculus” was used to divide students who had never taken calculus from those who had, “high school calculus” divided students who had taken high school calculus from those who had not taken high school calculus, which could include students who had not taken calculus and those students who had taken calculus in college. Because of this, we created six dichotomous variables, one for each possible mark on the questions, recognizing that there is a great deal of overlap between these choices. Each one of these six dichotomous variables then separates the entire population into two groups, and a t -test was run, predicting individual normalized gains for each of these variables, displayed in Table 10.

Table 11: t-test Results from Previous Mathematics Course Questionnaire

Variable	df	t -value	p -value
No Precalculus	11.553	-0.501	0.626
High School Precalculus	122.505	-1.904	0.059
College Precalculus	187.864	3.488	0.001
No Calculus	319.323	-0.816	0.415
High School Calculus	463.496	-1.045	0.297
College Calculus	87.162	2.517	0.014

The results of the table are that the response to having taken college precalculus and having taken college calculus were each significant predictors of CCI gains. Taking college precalculus as an example, we interpret the small p -value associated with the college precalculus variable to mean that the normalized gain scores of students who previously took college precalculus ($M = 0.180$, $SD = 0.238$) is significantly different

from those who did not ($M = 0.267$, $SD = 0.223$). The directionality of the comparison can be determined by either considering the means of each group to see that the mean of students who did not take college precalculus is higher than the mean of students who did, or by considering the sign of the t -value in the table. Since the t -value is positive, students who said that they had taken college precalculus would have lower scores than those who did not. Similarly, the small p -value associated with the question of having previously taken college calculus along with the positive sign of the t -value indicates that students who had previously taken college calculus had smaller gains than those who had not. This is reflected in the mean of the students who had previously taken college calculus ($M = 0.179$, $SD = 0.235$) and of the students who had not previously taken college calculus ($M = 0.256$, $SD = 0.227$). These results suggest that dividing students based on previous mathematics courses may be useful.

The six dichotomous variables used in the previous set of t -tests provides some initial information, but sometimes groups students in ways that are difficult to interpret. For example, students who reported that they had not previously taken precalculus in high school may represent students who had taken precalculus in college, as well as students who never took precalculus. For this reason, we decided to create a set of categories which exhaustively categorized all the students in the study into meaningful categories based on their prior mathematics course background. The categories created are:

1. Students who have previously taken calculus in college – these students are repeating the same course again, whether at the same or another university, termed “college repeaters” ($N = 67$).
2. Students who have not previously taken college calculus, but have taken high school calculus, termed “high school repeaters” ($N = 247$).
3. Students who have not taken calculus in high school or college, but have taken

- college precalculus, termed “first-time calculus, less prepared students” ($N = 59$).
4. Students who have not taken calculus in higher school or college, and have not taken college precalculus. The students in this category are those who might be considered to be strong enough mathematics students that they were able to take college calculus without needing to take college precalculus, but were not strong enough mathematics students in high school that they did not take calculus in high school, or were in high schools that did not offer calculus, termed “first-time calculus, better prepared students” ($N = 109$).

These categories were constructed because each population has unique traits that may help or hinder their abilities in college calculus, such as prior exposure to calculus concepts, and also because certain combinations of these groups identify additional populations of interest. For example, Groups 1 and 2 combined form all the students who have previously seen calculus, and Groups 3 and 4 together form all the students who have not taken calculus. Groups 1 and 3 consist of students who have previously taken a college level mathematics class. Students in Group 2 have not taken college calculus, but have taken high school calculus, so are mostly students who have not taken any college mathematics courses. There are 24 of the 247 students in Group 2 who took college precalculus, meaning that, after taking high school calculus, these students placed into college level precalculus or lower. This is important because if we want to separate students based on whether they have taken a previous college mathematics course, these 24 students will need to be moved into another group. Group 4 consists of students who have not taken college level calculus or precalculus. It is possible that these students have previously taken college mathematics courses other than precalculus or calculus, though we do not have access to this information. We will categorize the students in this category as having not taken a prior college level mathematics course, making the stipulation that the college mathematics courses of interest are precalculus and calculus. Except for the

24 students in Group 2, combining Group 1 with Group 3 and Group 2 with Group 4 divides the students into students who have taken college-level mathematics classes from those who have not. The mean normalized gains for each group are given in the table below.

Table 12: Normalized Gains for Students Based on Previous Mathematics Courses

Group	1	2	3	4
Group name	“college repeaters”	“high school repeaters”	“first-time calculus, less prepared students”	“first-time calculus, better prepared students”
Mean normalized gain	0.179	0.256	0.220	0.278

We initially test whether there are pairwise differences in their normalized gain scores. A pairwise t -test was conducted between the four groups, and the p -values of each comparison are shown in Table 13.

Table 13: p -values for Pairwise t -test Between Mathematics Background Groups

	Group 1	Group 2	Group 3
Group 2	0.08		
Group 3	0.83	0.83	
Group 4	0.03	0.83	0.45

This table demonstrates that the only significant difference between groups is between groups 1 and 4, meaning that there is a difference between students repeating the course and those who are taking the course for the first time, and have not previously taken any college level mathematics courses.

Groups 1 and 4 are different in two ways. Students in Group 1 have already seen

calculus before, and have also already taken a college level mathematics course. Students in Group 4 have not seen calculus, and have not taken a college level mathematics course (with the possible exception of a class other than precalculus or calculus). It is possible that the differences between these groups is attributable to one of these reasons, or both. By using different groupings of students, we can determine what role prior exposure to calculus and having prior experience in a college level mathematics class (precalculus or calculus) has on CCI gains.

First, we distinguish between students who have seen calculus before (Groups 1 and 2) from students who have not seen calculus before (Groups 3 and 4). A t -test reveals no difference between students who have seen calculus and those who have not seen calculus, $t(319.32) = -0.82, p = 0.42$. This is somewhat surprising since one might reasonably think that prior exposure to course content could either benefit students by allowing them to concentrate on the conceptual underpinnings of the subject since the procedural aspects would have been previously seen, or could be detrimental to students who would be less likely to focus on conceptual material, seeing it as building up to the procedures which they are already familiar with. By removing students who are repeating the same course, we see that comparing Group 2 with Groups 3 and 4 combined still does not result in a statistically significant difference, $t(333.67) = -0.086, p = 0.932$. Since anecdotal evidence suggests either of the above two scenarios, it is possible that both are occurring with similar frequency, effectively negating each other when compared with students who have not previously seen the content of the course.

We next distinguish between students who have taken a college mathematics class

before (Groups 1 and 3 with the previously mentioned 24 students added) from those who have not (Groups 2 and 4, with the previously mentioned 24 students removed). The 24 students are moved, as they have taken college precalculus. There is a significant difference between these two groups, $t(271.915) = -3.53$, $p < 0.001$. The mean normalized gain score for students who have taken a college mathematics class before is 0.190 ($SD = 0.237$) while the mean normalized gain score for students who have not taken a college mathematics course before is 0.271 ($SD = 0.222$). This has potentially important implications for college-level calculus instruction, though care must be taken in interpreting the results. While student self-selection and affective variables are very relevant to this discussion, this may be an indication that a student who is capable of taking calculus upon entering college may achieve greater gains by taking calculus immediately instead of taking precalculus first. Another interpretation might be that students who take college calculus immediately achieve higher gains because they have stronger math ability to start with, which might be one of the reasons they take college calculus without taking college precalculus. Before making any conclusive interpretation, however, further study would be needed to determine whether other factors would be more important in making this decision.

5.1 HLM Using Prior Mathematics Course Groupings

Using these groupings, we can revisit using an HLM to predict individual-level normalized gains. In this section, we predict individual gain scores ($N = 130$) using the divisions of students into groups defined as in the previous section as well as the measures of IE used in the analysis of gender.

We begin by considering encouraging revisions as a measure of IE along with the student grouping from the prior section in an HLM, the results of which are in Table 14 as Model 1. The results of using all interactions as a measure of IE instead of encouraging revisions are presented in the same table as Model 2. Since group membership is a categorical variable, there is a coefficient estimated for each group, except the baseline group, which is Group 1 by default. This choice makes no difference in the estimate of the coefficient for the IE variable, as demonstrated in Table 14 with Models 3 and 4 which create the same models as Model 1 and 2 respectively using no baseline group instead of Group 1. The estimates for the groups are different as they are now compared to 0 instead of each other, but in each case, the coefficient for the measure of IE can be interpreted as the impact of the IE measure on gain scores with group membership considered as an additional predictor variable in the model.

Table 14: HLM Predicting Individual Normalized Gains

<i>Fixed Effects</i>	Model 1	Model 2	Model 3	Model 4
Intercept (SE) t-value	0.008 0.087 0.091	-0.170 0.148 -1.146		
Group 1 (SE) t-value			0.008 0.087 0.091	-0.170 0.148 -1.146
Group 2 (SE) t-value	0.152 0.078 1.932	0.152 0.079 1.931	0.160 0.042 3.756	-0.018 0.119 -0.150
Group 3 (SE) t-value	0.224 0.104 2.153	0.225 0.104 2.156	0.232 0.073 3.160	0.055 0.125 0.441
Group 4 (SE) t-value	0.198 0.082 2.425	0.196 0.082 2.398	0.206 0.048 4.277	0.026 0.122 0.215
Number of revisions encouraged (SE) t-value	0.016 0.007 2.269		0.016 0.007 2.269	
All Interactions (SE) t-value		0.003 0.001 2.201		0.003 0.001 2.201
<i>Random Effects</i>				
Intercept Variance	< 0.001	< 0.001	< 0.001	< 0.001
Residual Variance	0.044	0.044	0.044	0.044
% Variance Explained	< 0.001	< 0.001	< 0.001	< 0.001

The results in this table indicate that once prior mathematics courses are taken into account, both the number of revisions encouraged and the number of all interactions are significant predictors of individual normalized gain scores, as demonstrated by the *t*-values for each of their coefficients (2.269 and 2.201 respectively).

In order to interpret these results, we first consider correlations between our predictor

variables. Our first concern is that the new variable of mathematics background grouping may be correlated with classrooms, suggesting that the classroom differences may actually be a result of self-sorting by the students. There is, however, not a significant difference in mathematics backgrounds among different instructors videotaped in the study, $\chi^2(12, N = 130) = 13.90, p = 0.31$, though there is a difference in the students at large, $\chi^2(66, N = 481) = 113.40, p < 0.001$. This suggests that when considering variables from the coding protocol, and hence only looking at students who were taught by instructors that were videotaped, there is no significant difference between classrooms in terms of mathematical background, as measured by prior courses. Since there is no difference between mathematics course backgrounds between the instructors, we interpret this to suggest that when mathematics course background is controlled for, students have higher gains when their instructors have more interactions of the “encouraging revisions” or “all interactions” types.

5.2 Possible Correlation with SAT Scores

One additional variable of interest is SAT mathematics scores, as it indicates a level of prior mathematics knowledge. There is a strong correlation between having taken a prior mathematics course and SAT mathematics scores, $r(178) = -0.274, p < 0.001$, which is not surprising since SAT mathematics scores are often used for placement. There is also a strong correlation between having taken a prior mathematics course and placement test scores, $r(423) = -0.283, p < 0.001$, again not surprising for the same reason as SAT mathematics scores. The placement scores are based on ALEKS, which is an online testing system based on learning space and knowledge space theory (Albert & Lukas,

1999; Doignon & Falmagne, 1999; Falmagne, Koppen, Villano, Doignon, & Johannesen, 1990), and has been previously utilized in the enhancement of college algebra instruction (Hagerty & Smith, 2005).

Each of the variables of ALEKS placement scores and having taken a prior college mathematics course are significant predictors of normalized gain scores, even when additional variables such as gender, year in college, and ethnicity are controlled for. An ANOVA indicates that there is a significant difference in ALEKS placement test scores between different mathematics course backgrounds, $F(3,421) = 8.103$, $p < 0.001$. None of the additional variables significantly predict normalized gain scores when either ALEKS scores or prior college mathematics courses are included in the model. When including both placement scores and prior college mathematics class in the model, both variables are statically significant, as seen in Table 15 by the p -values for placement test scores and having taken a prior college mathematics course being less than 0.05.

Table 15: Predicting Gains Based on Prior Mathematics Background and Placement Test
Dependent Variable: Normalized Gains

Variable	B	SE(B)	β	t	Sig (p)
Intercept	0.112	0.059		1.91	0.057
Placement test score	0.002	0.001	0.136	2.73	0.007*
Having taken a prior college mathematics course	-0.063	0.026	-0.122	-2.46	0.014*

Note: $R^2 = 0.04264$ (Adjusted $R^2 = 0.03811$). B indicates unstandardized regression coefficient. β indicates standardized regression coefficient.

* $p < 0.05$

The coefficient for placement test score is positive, indicating that having a higher

placement test score is positively correlated with higher normalized gain scores. The coefficient for having taken a prior college mathematics course is negative, meaning that taking a college math course is negatively correlated with normalized gain scores.

Students who have achieved higher gain scores may have done so due to higher mathematical ability level, as measured by their placement test scores. The metric used here may also be affecting the results we find. While we might expect a negative correlation between pretest scores and gain scores since a student with a greater mathematical ability level would have less room for improvement, the choice of normalized gain scores allows students with high pretest scores to achieve high gain scores. The model indicates that students with greater incoming mathematical knowledge as measured by the placement test are achieving greater gains than those students with weaker mathematical ability. The possible connection between gain scores and mathematical ability is also seen in the result that students who have previously taken a college mathematics course have lower gain scores than those who have not. The students who have taken mathematics courses before calculus at the college level may have weaker mathematical ability, leading to the negative correlation between having taken a college mathematics course and normalized gain scores.

It is likely that many factors such as student confidence play a large role in comparing students with different backgrounds and placement test scores. If two students enter a university with the same placement test score, and one enters calculus and the other precalculus, there are likely many differences in their attitudes towards mathematics. These differences may be leveraged so that each student population can be helped most

effectively.

6 Conclusions

This chapter further develops the analysis of the previous chapter by analyzing gain scores at the individual level. While the initial motivation for creating an HLM was to consider the effect that individual-level variables such as gender may have on CCI gains, additional results became apparent. Initially, performing analysis of the scores reported in Chapter 2 using a null model with individual-level normalized gains indicated that the relationships between IE level and student gains may have been a spurious relationship in our data. The difference in results between an instructor-level model with average gains and a HLM with individual student gains clustered by instructor suggests that more investigation needs to be conducted. By using the coding protocol to quantify levels of IE instead of binning classes and following a similar method of analysis as previous studies using concept inventories, it appeared that increasing the degree of types of IE in a classroom positively affected student learning. When the HLM was initially run, however, the variance of individual normalized gain scores was accounted for, and this relationship was no longer present.

While analysis of gender revealed that a gender gap did exist in our data, the gap did not significantly change over the course of the semester. Gender also seemed to have no relationship with IE instructional styles in encouraging CCI gains. While results have shown interactive teaching techniques to decrease the gender gap in mathematics classrooms (Laursen et al., 2011), this has not been investigated using the particular framework of IE. The IE framework has, however, been investigated along with gender in

demonstrating a reduced gender gap on concept inventories in physics (Kost et al., 2009), suggesting that a relationship may exist in some form in mathematics.

When considering previous mathematics experience, there is some evidence that IE instructional styles may encourage CCI gains. Only when this variable is included in the HLM do the two IE variables which predicted CCI gains at the classroom level become significant again. When the variance associated with prior mathematics experience is removed, this seems to create between instructor variance, which can then be explained by encouraging revisions. This may be due to an additional relationship between a student's mathematics background and the effect that IE instructional techniques have on their learning. While this study only establishes initial evidence for this claim, it is a promising area for future research to be conducted.

CHAPTER 4: COMPARING GAIN SCORE MEASURES ON THE CCI

In Chapters 2 and 3, we used normalized gain scores to measure class and student improvements on the CCI. In this chapter, we introduce another method of measuring gains. The framework for creating these scores is called Item Response Theory (IRT). Item response theory is not frequently used in concept inventory studies (Wallace & Bailey, 2010), though it has been used to study the FCI (Morris et al., 2012; Wang & Bao, 2010), concept inventories in astronomy (Aslanides & Savage, 2013; Favia, Comins, & Thorpe, 2012; Schlingman, Prather, & Wallace, 2012; Wallace & Bailey, 2010), and statistics (Allen, 2007). An IRT analysis of the CCI has not been published, nor have multiple measures of gain been studied for this instrument. Our study contributes to the existing literature in both of these areas.

1 Introduction

Item Response Theory (IRT) is a modern approach to analyzing instruments like tests or surveys (Embretson & Reise, 2000). IRT is based on the idea that an instrument measures a latent trait or ability, such as conceptual knowledge of calculus in the case of the CCI. While this trait cannot be directly observed, the effects can be observed through answers to questions. Mislevy (n.d.) pointed out in an email to Hake that IRT has some benefits over the use of normalized gains such as handling floor and ceiling effects. For example, students who obtain a perfect score obtain a normalized gain of 1 regardless of initial ability levels. For these students, then, improving from a very low score to the maximum or a very high score to the maximum are not differentiated because the student has achieved the highest level possible on the test. Another benefit of IRT is that it allows

for the analysis of individual questions as opposed to a single test score for each individual as is used by normalized gains.

When creating an item response theory model, one first estimates parameters of the items on the instrument. Once these parameters are estimated, they can be used to estimate ability levels for each participant who completed the instrument. When the instrument is given as a pretest and a posttest, as was done with the CCI, the difference between the ability measured is the change over the course, and so it measures gain (Wallace & Bailey, 2010).

IRT can be used with many types of instruments. For example, IRT can be used to analyze responses to Likert scale questions which may address variables such as political inclination. For this study, all items are either considered correct or incorrect. In other assessment contexts, such as measuring political affiliation, items might indicate endorsement or lack of endorsement.

1.1 Relevance to Research Questions

This chapter directly addresses research question 6, “Do different methods for scoring conceptual knowledge gains correlate similarly with IE characteristics? How might we explain apparent differences and how important are they?”. In this chapter we discuss how IRT scores are constructed and how the IRT construction of scores differs from normalized gain scores. In the following chapter, we use the IRT gain scores as an alternative measure of gains on the CCI to further investigate the connection between IE instruction and CCI gains.

2 Results of IRT Analysis and Implications

2.1 Types of IRT Models

In this section, we describe the results of constructing different types of IRT models. While each model provides information which can be used to understand the instrument, statistical tests can also be run to determine which of the models is the best fit for each of the pretest and the posttest. These comparisons are made in Section 2.2.

2.1.1 Rasch Model

Item Response Theory (IRT) consists of a family of probabilistic models for estimating parameters of multiple choice test items and of those individuals taking the test. IRT posits the existence of a latent trait (or latent traits in the case of a multidimensional model) which is required to successfully answer the items on the instrument. In analyzing the CCI, we are then assuming there exists a trait which we might call “conceptual knowledge of calculus,” the evidence of which manifests in the ability to answer conceptual calculus questions correctly. The greater the ability of an individual, the greater the probability of answering the question correctly.

The Rasch model is the simplest of the IRT models, estimating a single parameter, called difficulty, for each test item. In the Rasch model, the probability of an individual, p , with ability level θ_p , correctly answering question i , which has difficulty β_i , is given by

$$P(X_{pi}=1|\theta_p, \beta_i) = \frac{\exp[\theta_p - \beta_i]}{1 + \exp[\theta_p - \beta_i]} \quad (4)$$

For each specific item i and each person p , once the difficulty of that item and ability of that person have been estimated, a probability that the person will answer the question correctly can be computed, which is dependent upon the difference between the person's

ability and the item's difficulty. For an individual item, one can plot the probability of answering correctly against ability level, resulting in a logistic curve called the item characteristic curve (Embretson & Reise, 2000). When Rasch models of the CCI pretest and CCI posttest are created, a family of curves is created for each item and each test. The item characteristic curves for the CCI pretest as shown in Figure 7 and for the CCI posttest in Figure 8. In the pretest, items are labeled “aq” followed by the number of the test item. On the posttest, items are labeled “cq” followed by the number of the test item.

Figure 7: Rasch Model for CCI Pretest

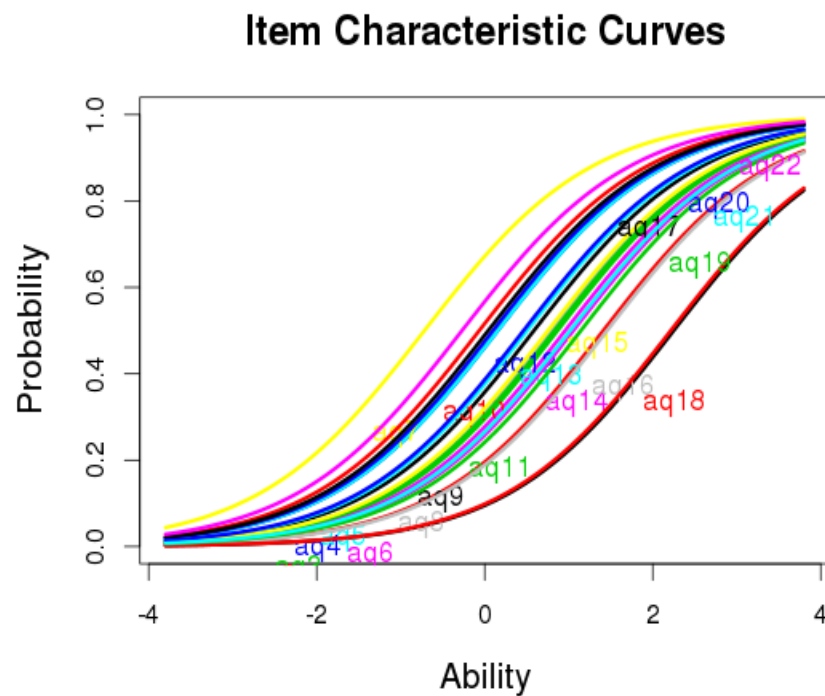
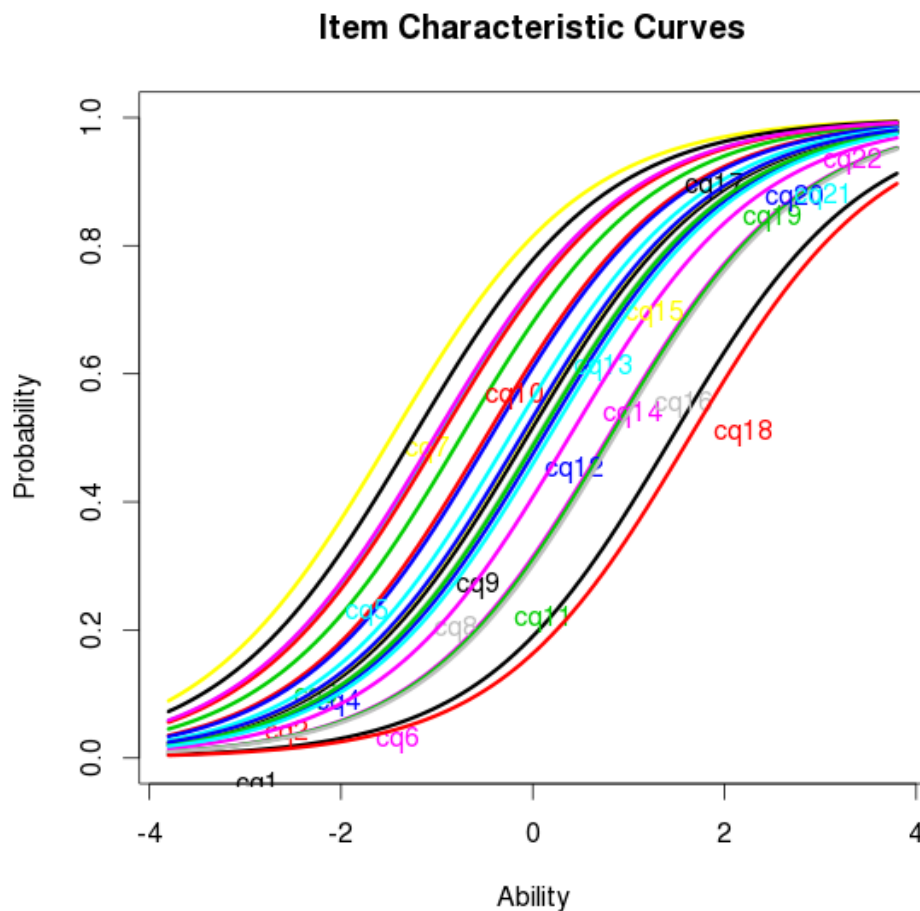


Figure 8: Rasch Model for CCI Posttest



Given these graphs, we can interpret from either an item-based or person-based perspective. For an example of an item-based perspective, we can consider item 18 on the pretest, which is the red curve in Figure 8 above, farthest to the right. On this particular item, the probability of answering correctly increases with ability, suggesting that individuals with a high ability level are more likely to answer the question correctly. This ability is the latent trait which we assume the instrument to be measuring, in this case, conceptual knowledge in calculus. For this particular item, the ability needed to achieve a

50% probability of answering the item correctly is roughly two (the exact values are given in Table 16), meaning that an individual who is about two standard deviations above the mean ability level has a roughly 50% chance of answering the question correctly. Individuals with a higher ability level than this are more likely to answer the question correctly, and individuals with a lower ability level are less likely to answer the question correctly. This makes item 18 the most difficult item on the posttest. We can also interpret this graph from the perspective of an individual test-taker. Each person is assumed to have a single ability level, θ_p . For this ability level, a corresponding probability is found for each item, indicating the likelihood of that person answering that particular item correctly. One of the features of the Rasch model not shared by all IRT models is that the order of difficulty of the items is independent of the individual taking the test (Wallace & Bailey, 2010). These difficulty values indicate the ability level required to achieve a 50% chance of answering that item correctly.

Table 16: Rasch Model Estimated Parameters

	Pretest difficulties (β_i)	Posttest difficulties (β_i)
Question 1	2.25	1.45
Question 2	1.41	-0.50
Question 3	0.87	-0.75
Question 4	0.16	-0.45
Question 5	0.15	-1.01
Question 6	0.96	0.78
Question 7	-0.72	-1.48
Question 8	0.79	0.04
Question 9	0.59	-0.07
Question 10	-0.06	-0.97
Question 11	0.83	0.80
Question 12	0.10	0.10
Question 13	0.51	-0.25
Question 14	1.06	0.37
Question 15	0.75	-0.11
Question 16	1.46	0.84
Question 17	0.04	-1.25
Question 18	2.21	1.63
Question 19	1.15	0.05
Question 20	0.46	0.13
Question 21	1.02	0.16
Question 22	-0.27	-1.03

The ordering of the difficulty of the items does not seem to change drastically between the pretest and the posttest. For example, the easiest item on the pretest is question 7 and is also the easiest question on the posttest. Furthermore, the signs of the difficulties do not change frequently. Different software for constructing IRT models normalize scores in different ways, as one can either normalize the item difficulties around 0 or normalize

individual ability levels around 0. Once either difficulties or abilities are centered around 0, the other set of parameters is fixed. This leads to slightly different interpretations of a “0 ability” in the model. If the software centers around individual ability, 0 means a student whose ability is average. If the software centers around item difficulties, a score of 0 means that an individual performs at the average difficulty of the test. The ltm (latent trait modeling) package in R was used to construct these scores centers around ability level, so the difficulties in the table above are in relation to an average student taking that particular test. While one can convert between the different methods of normalizing by a linear transformation, we kept the data in this form for ease of interpretation. For example, the difficulty of item 8 on the pretest was 0.79 while it was 0.04 on the posttest. This means that an average student was more likely to solve the question correctly on the posttest than on the pretest, or that the item appeared easier to those students taking the posttest than to those students taking the pretest. No items appeared more difficult on the posttest than the pretest, but item 12 had the same difficulty on the pretest as the posttest, and the difficulty of item 11 changed very little. This indicates that the content of these two items does not seem to be addressed during the course in such a way that students are able to more easily answer the same question at the end of the course as the beginning.

2.1.2 One Parameter Logistic Model

One of the restrictions of the Rasch model is that it imposes a strict relationship between the change in ability and change in probability of correctly answering a question. To address this potential concern, one can construct a model called a one parameter logistic (1PL) model by introducing a new parameter, α , called the discrimination,

estimated from the data. The Rasch model can be thought of as a special case of the 1PL model in which the discrimination parameter is forced to be 1. The effect of the discrimination parameter is to identically change the slopes of all the item characteristic curves, while interpretations of θ_p and β_i remain the same as in the Rasch model. This model is given by the formula:

$$P(X_{pi}=1|\alpha, \theta_p, \beta_i) = \frac{\exp[\alpha(\theta_p - \beta_i)]}{(1 + \exp[\alpha(\theta_p - \beta_i)])} \quad (5)$$

The larger the value of α , the steeper the slopes of the curves, and the more discriminating the items. The effect of the discrimination parameter might be best understood by the two most extreme cases. Considering a single item, a discrimination near 0 would produce a nearly flat curve, suggesting that the probability of correctly answering the item does not change much as ability changes. A very large discrimination would produce a characteristic curve near a step function, suggesting that any ability (θ_p) above the difficulty of the item (β_i) would nearly guarantee getting the item correct while an ability less than the difficulty of the item would nearly guarantee answering the question incorrectly. A plot of the item characteristic curves for the CCI pretest modeled with a 1PL model is given in Figure 9 and the 1PL CCI posttest model in Figure 10. The parameters for the 1PL models, which consists of difficulties for each item and a single discrimination parameter for each test are given in Table 17.

Figure 9: 1PL Model for CCI Pretest

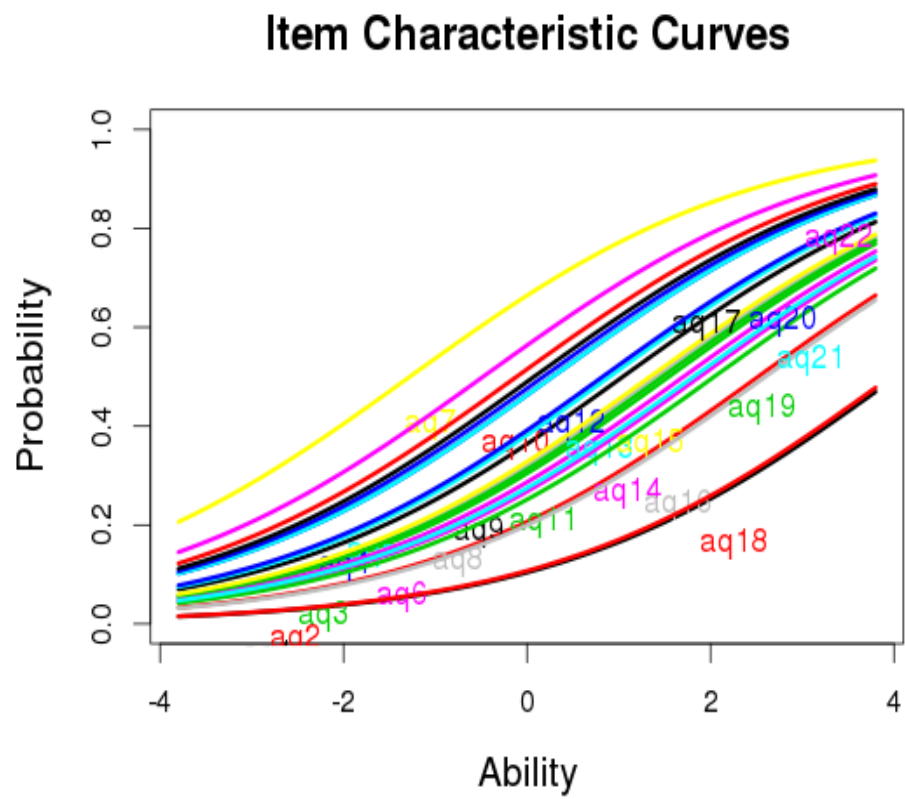


Figure 10: IPL Model for CCI Posttest

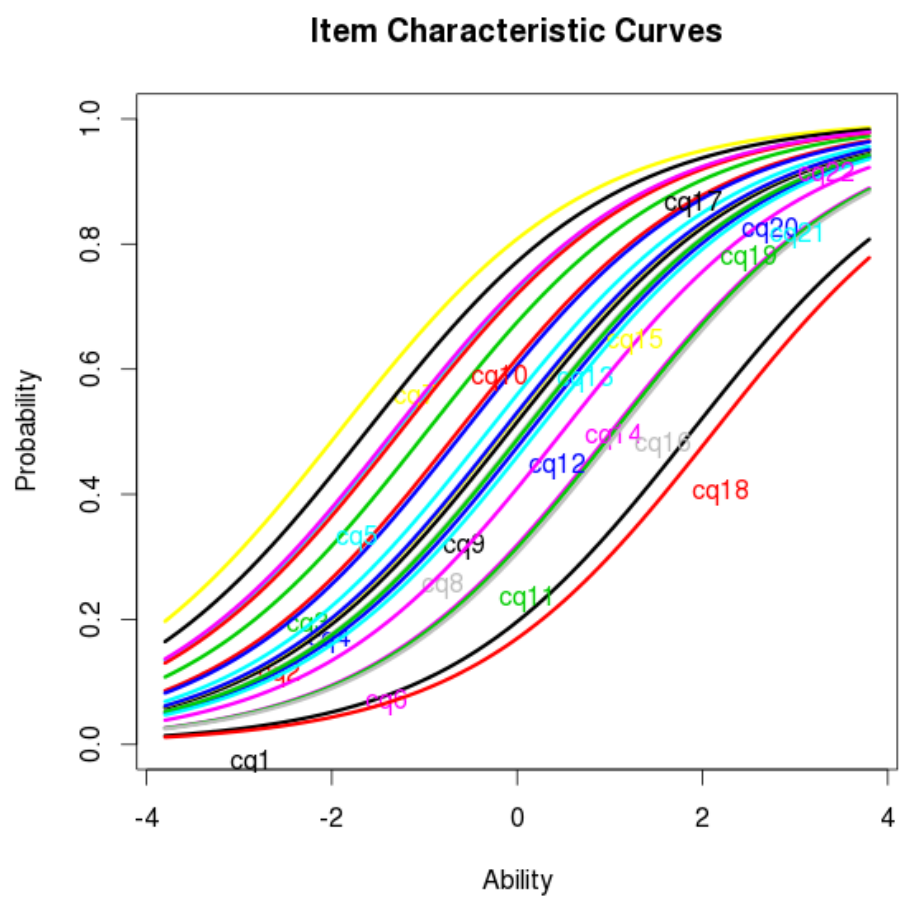


Table 17: 1PL Model Estimated Parameters

	Pretest		Posttest	
	Difficulty (β_i)	Discrimination (α)	Difficulty (β_i)	Discrimination (α)
Question 1	4.03	0.53	1.88	0.75
Question 2	2.52	0.53	-0.64	0.75
Question 3	1.55	0.53	-0.98	0.75
Question 4	0.28	0.53	-0.58	0.75
Question 5	0.28	0.53	-1.32	0.75
Question 6	1.71	0.53	1.01	0.75
Question 7	-1.28	0.53	-1.92	0.75
Question 8	1.40	0.53	0.05	0.75
Question 9	1.04	0.53	-0.08	0.75
Question 10	-0.11	0.53	-1.26	0.75
Question 11	1.47	0.53	1.04	0.75
Question 12	0.18	0.53	0.13	0.75
Question 13	0.90	0.53	-0.32	0.75
Question 14	1.88	0.53	0.49	0.75
Question 15	1.34	0.53	-0.15	0.75
Question 16	2.60	0.53	1.09	0.75
Question 17	0.08	0.53	-1.63	0.75
Question 18	3.96	0.53	2.12	0.75
Question 19	2.04	0.53	0.07	0.75
Question 20	0.82	0.53	-0.16	0.75
Question 21	1.82	0.53	0.21	0.75
Question 22	-0.48	0.53	-1.34	0.75

We find similar patterns in the 1PL model as the Rasch model. Questions 1 and 18 are still the most difficult items, and questions are less difficult for students on the posttest than on the pretest. We also see that the discrimination parameters on both tests are less than 1. Since the 1PL model with $\alpha = 1$ would result in the Rasch model, the 1PL model

suggests that the items are less discriminating than would be estimated using a Rasch model. The discrimination parameter is larger on the posttest than the pretest, suggesting that the questions better differentiate ability levels on the posttest than on the pretest.

2.1.3 Two Parameter Logistic Model

The difference between the Rasch and 1PL models is the addition of the discrimination parameter. The 1PL model, however, requires that the discrimination parameter must be the same for all items on a test. A two parameter logistic model (2PL) relaxes the condition that the discrimination parameter must be the same for all items, so that a difficulty and discrimination will be estimated for each item on the test. This model can be written as:

$$P(X_{pi}=1|\alpha_i, \theta_p, \beta_i) = \frac{\exp[\alpha_i(\theta_p - \beta_i)]}{1 + \exp[\alpha_i(\theta_p - \beta_i)]} \quad (6)$$

Graphs of the item characteristic curves in the 2PL model of the CCI pretest and posttest are given in Figure 11 and Figure 12, and estimated parameters are given in Table 18.

Figure 11: 2PL Model for CCI Pretest

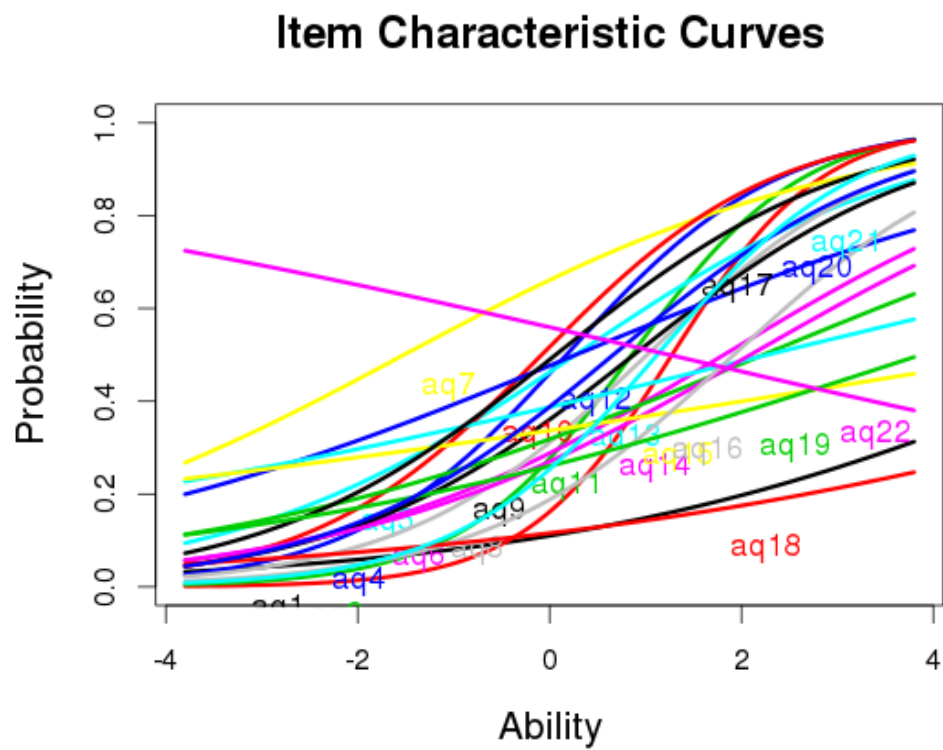


Figure 12: 2PL Model for CCI Posttest

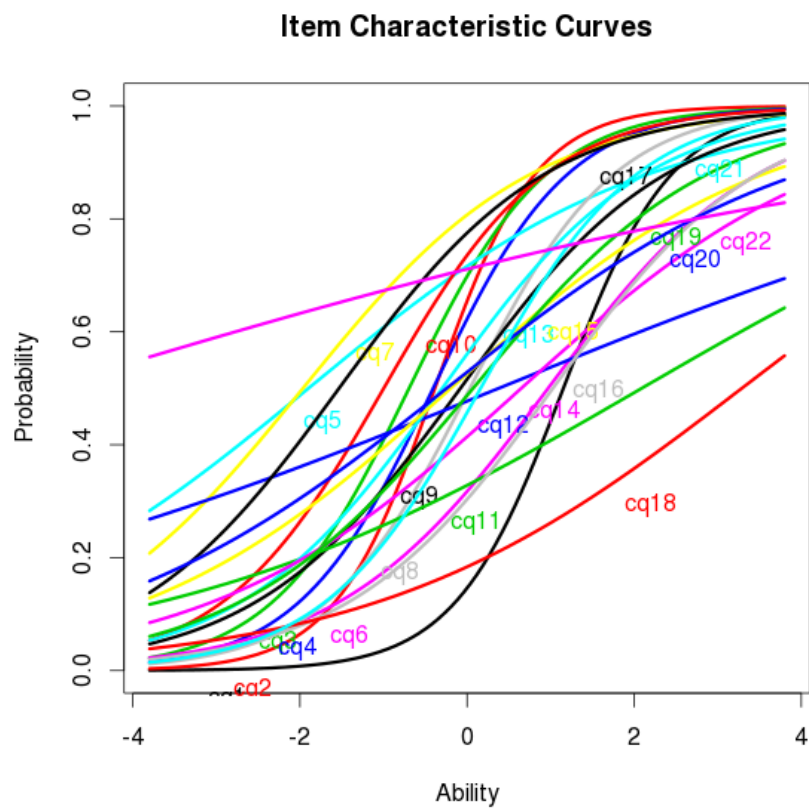


Table 18: 2PL Model Estimated Parameters

	Pretest		Posttest	
	Difficulty (β_i)	Discrimination (α_i)	Difficulty (β_i)	Discrimination (α_i)
Question 1	6.09	0.34	1.15	1.53
Question 2	1.28	1.29	-0.39	1.66
Question 3	0.86	1.12	-0.69	1.21
Question 4	0.18	0.91	-0.41	1.23
Question 5	0.27	0.56	-1.90	0.49
Question 6	1.82	0.50	0.97	0.79
Question 7	-1.52	0.44	-1.96	0.73
Question 8	1.02	0.78	0.03	1.15
Question 9	0.88	0.65	-0.08	0.81
Question 10	-0.08	0.83	-1.00	1.03
Question 11	2.23	0.34	2.09	0.34
Question 12	0.27	0.34	0.38	0.24
Question 13	2.27	0.20	-0.30	0.82
Question 14	2.09	0.47	0.65	0.53
Question 15	5.00	0.14	-0.19	0.53
Question 16	1.92	0.76	1.03	0.81
Question 17	0.06	0.66	-1.53	0.81
Question 18	8.41	0.24	3.29	0.45
Question 19	3.87	0.27	0.07	0.71
Question 20	0.66	0.69	-0.24	0.47
Question 21	1.12	0.96	0.16	1.06
Question 22	1.25	-0.19	-5.05	0.18

Inspecting Figure 11, it is apparent that the behavior of the last item on the test is counter to what one would expect for a test measuring a single construct: the item characteristic curve for item 22 is decreasing, indicating that as one's conceptual knowledge of calculus increases, the likelihood of answering the item correctly

decreases. This particular behavior would not be observed in the Rasch model since the item characteristic curves in a Rasch model are increasing functions. Whether the item characteristic curves are increasing or decreasing is a function of whether the discrimination parameter is positive or negative. In a 1PL model, all items have the same value of α , so are either all increasing or all decreasing. We expect the functions to be increasing since we expect students with higher ability levels to have a higher probability of answering each question correctly. It is not until the 2PL model, however, that the parameters are free enough to have a single item's item characteristic curve which is decreasing.

An initial suspicion was that test fatigue may have factored into the unexpected behavior of item 22 since it was the final item on the pretest. Students may have felt pressured for time or been tired by the end of the test and were simply guessing at that point. Since the items were reordered when giving the pretest and the posttest, item 22 was the last item on the pretest, but was numbered 20 on the posttest. On the posttest, the same item stands out as being particularly easy with a difficulty of -5.05 and does not discriminate well with a discrimination of 0.18. Taken together, this suggests that many students answered the question correctly, so it is not surprising that the item is unable to distinguish individuals very effectively. The item is reasonably difficult on the pretest, with a difficulty of 1.25. Inspection of the item provided no insight into the behavior of the item, so the item was removed from future analysis, as is frequently done in concept inventory studies utilizing IRT (Wallace & Bailey, 2010).

2.1.4 Three Parameter Logistic Model

The final model which will be explored here, introduces an additional parameter that accounts for the possibility of guessing, called the three parameter logistic (3PL) model. Since IRT is used to analyze multiple choice instruments, it is possible that on a test of skill individuals with low ability level are simply guessing. In a Rasch, 1PL, or 2PL model, the probability of answering a question correctly as ability decreases approaches zero. In a 3PL model, the new parameter, denoted γ_i , represents the lower limit of probability of answering a question correctly. The model is given by the formula:

$$P(X_{pi} = 1 | \alpha_i, \theta_p, \beta_i) = (1 - \gamma_i) \frac{\exp[\alpha_i(\theta_p - \beta_i)]}{1 + \exp[\alpha_i(\theta_p - \beta_i)]} + \gamma_i \quad (7)$$

The value of the guessing parameter γ_i is, like all item parameters, determined completely by the data. One might expect that a question with five possible choices would result in a guessing parameter of 0.2 since an individual with no knowledge of the topic would guess randomly and have a one in five chance of answering correctly. Different questions may have incorrect answer choices with different effectiveness, for example some answer choices may be ruled out by students. A different guessing parameter is then estimated for each item. Graphs of the 3PL models of the CCI pretest and posttest are given in Figure 13 and Figure 14, and a table of estimated parameters is given in Table 19.

Figure 13: 3PL Model for CCI Pretest

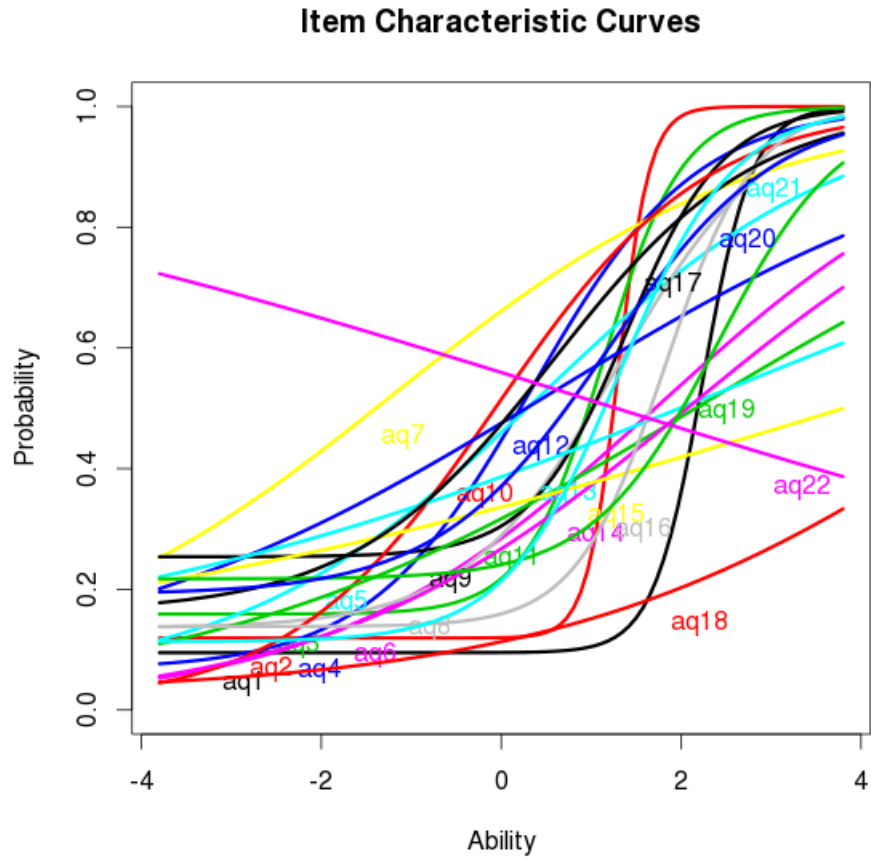


Figure 14: 3PL Model for CCI Posttest

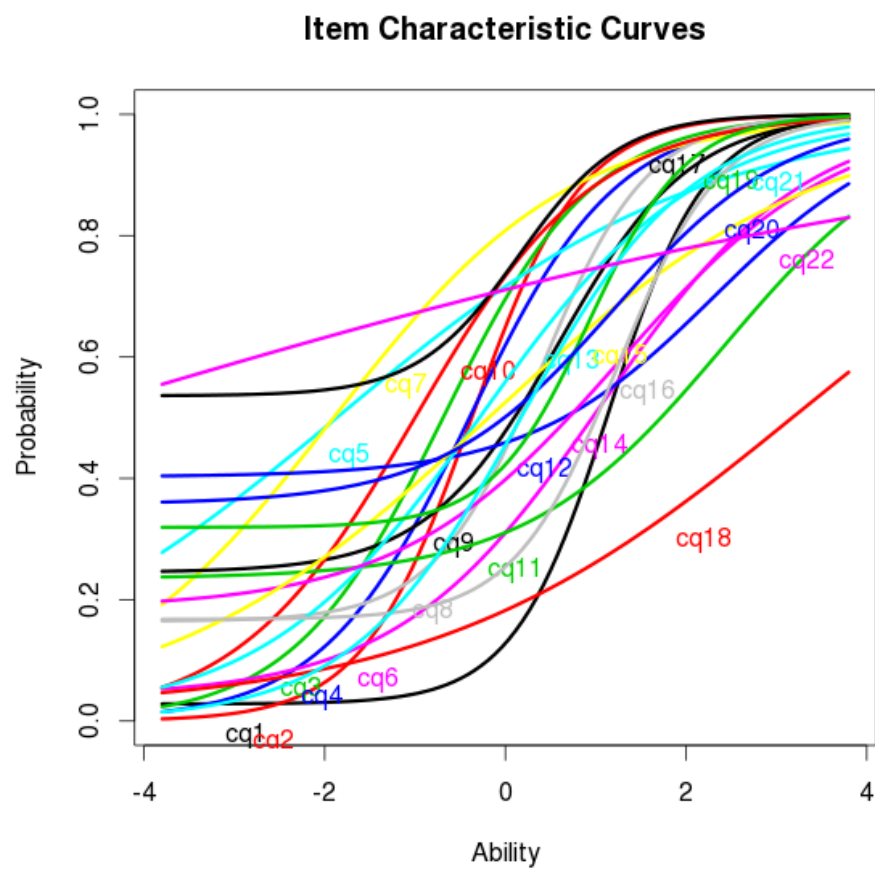


Table 19: 3PL Model Estimated Parameters

	Pretest			Posttest		
	Guessing	Difficulty	Discrimination	Guessing	Difficulty	Discrimination
	(γ_i)	(β_i)	(α_i)	(γ_i)	(β_i)	(α_i)
Question 1	0.10	2.24	3.63	0.03	1.15	1.88
Question 2	0.12	1.30	5.69	0.00	-0.37	1.65
Question 3	0.16	1.14	2.28	0.00	-0.69	1.20
Question 4	0.07	0.34	1.10	0.00	-0.40	1.23
Question 5	0.04	0.45	0.59	0.00	-1.85	0.50
Question 6	0.01	1.73	0.54	0.04	1.06	0.89
Question 7	0.05	-1.19	0.50	0.00	-1.90	0.76
Question 8	0.14	1.29	1.19	0.16	0.40	1.68
Question 9	0.25	1.39	1.87	0.25	0.59	1.38
Question 10	0.01	-0.05	0.86	0.00	-1.00	1.01
Question 11	0.01	2.19	0.36	0.23	2.42	0.91
Question 12	0.05	0.53	0.38	0.40	2.32	0.97
Question 13	0.07	2.56	0.26	0.01	-0.27	0.83
Question 14	0.00	2.08	0.49	0.19	1.26	0.83
Question 15	0.04	4.27	0.19	0.00	-0.17	0.55
Question 16	0.14	1.82	2.03	0.17	1.24	1.73
Question 17	0.16	0.58	0.90	0.54	0.15	1.78
Question 18	0.03	5.70	0.41	0.02	3.25	0.49
Question 19	0.22	2.41	1.44	0.32	0.93	1.85
Question 20	0.19	1.18	1.07	0.36	1.20	1.04
Question 21	0.11	1.31	1.57	0.00	0.17	1.06
Question 22	0.05	0.77	-0.20	0.03	-4.67	0.18

We see two items, questions 11 and 12, which appear more difficult on the posttest than the pretest, though all other items decrease in difficulty from the pretest to the posttest. Question 22 is still a poorly fitting item, with a negative discrimination

parameter on the pretest, and poor discrimination with very little difficulty on the posttest. Question 18 is still the most difficult question on both the pretest and the posttest. The low guessing parameter on the final item, which was of concern earlier, indicates that our initial hypothesis of test exhaustion or time pressure are not likely to be accurate, as it seems students were not guessing on these questions.

Some of the guessing parameters for other questions were particularly interesting. On the pretest, items 9, 19, and 20 all had relatively large guessing parameters. While items 19 and 20 included vocabulary which may have been unfamiliar to students with no exposure to calculus, question 9 which had the highest guessing parameter had no such language. On the posttest, items 9, 11, 12, 14, 17, 19, and 20 all had fairly large guessing parameters. Questions 9, 19, and 20 aside, question 11 included no explicit calculus knowledge, only knowledge of real numbers and the idea of limits, 12 required a notion of velocity and acceleration, and 14 and 17 both required explicit calculus knowledge. There does not seem to be any clear pattern, as there were many questions with low guessing parameters which required the same types of knowledge as questions with high guessing parameters. Student interviews may provide some greater insight into possible causes for these response patterns.

2.2 Comparing Models

The four models described above are nested in that each can be reduced to a prior model by making certain restrictions. A 3PL model in which the guessing parameter is forced to be zero is identical to a 2PL model. If the discrimination parameters in the 2PL model are all forced to be identical, this model is identical to a 1PL model. If the 1PL

model has a discrimination forced to be 1, the model reduces to a Rasch model. While adding additional parameters will fit the data better, we are also interested in finding parsimonious models. Just as nested regression models can be tested to determine whether the additional predictor variables significantly reduces errors, nested IRT models can also be tested to see whether the reduction of error is statistically significant given the additional parameters estimated.

The approach of testing nested IRT models to determine which model to use is not universally accepted (Andrich, 2004). Some believe that a model should be picked based on the measurement properties instead of by statistically testing the models, which is appealing particularly to those who prefer Rasch models. The measurement properties of Rasch models which make them appealing over other models will not be discussed here, but they are described by Wallace and Bailey (2010). We instead follow the approach of letting the data determine which model fits best by directly comparing the fit of the different models, and determining whether adding additional parameters reduces errors significantly. We pick this because we find that that the information provided by comparing the models is valuable.

Before performing the tests comparing the models, we consider what might be expected on an instrument like the CCI. The distractors on the CCI are specifically designed to capture specific misconceptions that students have about calculus (Epstein, 2007). For this reason, students who have a poor understanding of the fundamental concepts of calculus might be more likely to be drawn to specific incorrect answers (depending on the particular misunderstanding that they have) rather than just guess

randomly. It would be reasonable to suspect that the 2PL model would be a statistically better fit to the data than the 3PL model. Whether the Rasch, 1PL model, or 2PL model would fit best would largely be an issue of the specific questions themselves and how they behave relative to each other, so there is no way to guess which would fit the data best. For example, whether the 1PL or 2PL model fits better is a matter of how similar the discrimination parameters are to each other, which cannot be guessed on how the instrument was constructed. Determining which of the Rasch, 1PL, and 2PL models fits the data best must be determined from the data themselves. Given the variety of difficulties and discriminations we found in the previous analysis, we expect that the 2PL or 3PL will fit the data best, though we will test all models for completeness.

We will empirically test the pretest first, and will then test the posttest separately. In order to find the most parsimonious model, we compare models with increasing complexity. We begin by comparing the Rasch model of the pretest with the 1PL model of the pretest, which results in a chi-square test with one degree of freedom, as one new parameter has been introduced in the 1PL model over the Rasch model. The p -value for the test indicates whether the more complex model is a significantly better fit for the data, where a p -value less than 0.05 indicates that the more complex model is a better fit.

Table 20: Comparison of Pretest Rasch and Pretest 1PL Models

Likelihood Ratio Table

	AIC	BIC	Log Likelihood	LRT	df	Sig (p)
Rasch	23589.79	23694.95	-11772.89			
1PL	23356.60	23466.54	-11655.30	235.19	1	<0.001

We find that the introduction of the α parameter improves the fit of the model, as Table

20 indicates that the 1PL model is a better fit than the Rasch model for the CCI pretest.

We then compare the 1PL model with the 2PL model.

Table 21: Comparison of the Pretest 1PL Model and the Pretest 2PL Model
Likelihood Ratio Table

	AIC	BIC	Log Likelihood	LRT	df	Sig (p)
1PL	23356.60	23466.54	-11655.30			
2PL	23194.43	23404.74	-11553.21	204.17	21	<0.001

The small p -value in Table 21 indicates that the more complex model, the 2PL model, is a significantly better fit than the 1PL model.

We then compare the 3PL model to the 2PL model.

Table 22: Comparison of the Pretest 2PL Model and Pretest 3PL Model
Likelihood Ratio Table

	AIC	BIC	Log Likelihood	LRT	df	Sig (p)
2PL	23194.43	23404.74	-11553.21			
3PL	23184.63	23500.10	-11526.32	53.8	22	<0.001

Again, the p -value is less than 0.05, indicating that the more complex model, the 3PL model, is the best fit for the CCI pretest.

We now consider which model best describes the CCI posttest. We compare the Rasch and 1PL models, resulting in Table 23.

Table 23: Comparison of the Posttest Rasch Model and Posttest 1PL Model
Likelihood Ratio Table

	AIC	BIC	Log Likelihood	LRT	df	Sig (p)
Rasch	18113.87	18212.97	-9034.94			
1PL	18066.15	18169.75	-9010.07	49.73	1	<0.001

The small p -value indicates that the more complex model, the 1PL model, is the better fit.

We then compare the 1PL and 2PL models.

Table 24: Comparison of the Posttest 1PL Model to the Posttest 2PL Model

Likelihood Ratio Table						
	AIC	BIC	Log Likelihood	LRT	df	Sig (p)
1PL	18066.15	18169.75	-9010.07			
2PL	17915.78	18113.97	-8913.89	192.37	21	<0.001

These results again indicate that the more complex model, the 2PL model, is the better fit.

We finally compare the 2PL and 3PL models.

Table 25: Comparison of the Posttest 2PL Model with the Posttest 3PL Model

Likelihood Ratio Table						
	AIC	BIC	Log Likelihood	LRT	df	Sig (p)
2PL	17915.78	18113.97	-8913.89			
3PL	17937.14	18234.43	-8902.57	22.63	22	0.423

The p -value of 0.423 indicates that the 3PL model is not a significantly better fit than the 2PL model, so the 2PL model remains the best fitting model for the CCI posttest.

The difference between the pretest and posttest best-fitting models suggests that at the beginning of the semester, students with poor understanding of calculus concepts are guessing randomly, while students at the end of the semester who have poor understanding are being drawn to specific incorrect answers. Additional analysis of the data by modeling sub-populations of the students depending on whether they have taken calculus classes before may provide some additional information on this question, and student interviews would provide further information to answer this question. The CCI test could also be given again with a sub-question to each question asking how confident

the student is in his or her answer, directly addressing whether students are guessing or not.

There are a few possible reasons students may be guessing on the pretest but not on the posttest. One is that since some of the notation or vocabulary may be unfamiliar to students beginning calculus who have never seen it, students who do not understand the questions may be simply guessing. Another possible explanation is that the specific types of misconceptions about calculus develop while learning calculus. It is also possible that both of these reasons are affecting the propensity of students to randomly guess.

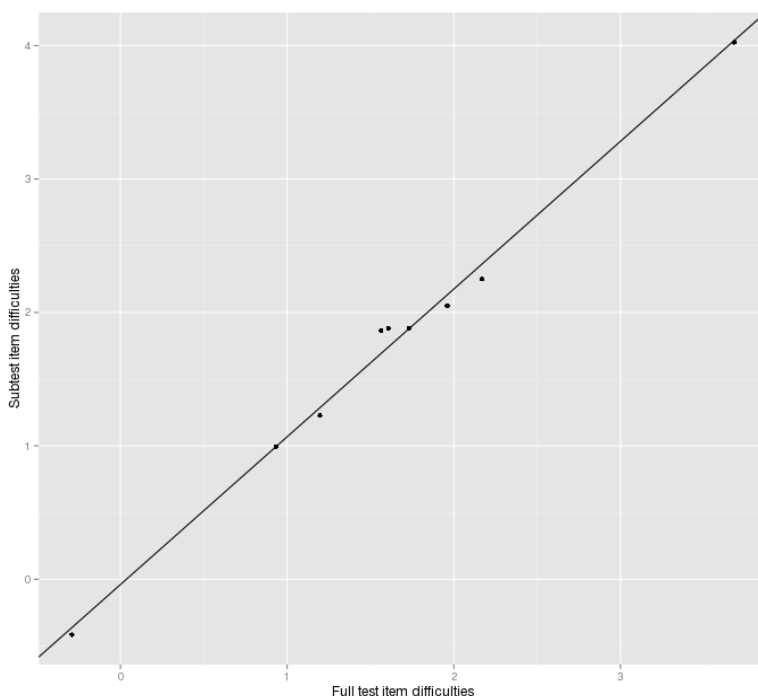
2.3 Checking Assumptions

There are two assumptions being made when constructing an IRT model: unidimensionality and local independence (Embretson & Reise, 2000). Unidimensionality ensures that the probabilities estimated are the result of a single construct or trait. Unidimensionality may be violated if multiple skills are required to solve a problem, such as mathematics word problems requiring both verbal skills to understand the problem and mathematical skills to solve the problem. Local independence means that no additional factors explain correlations in item responses (Embretson & Reise, 2000). Each of these assumptions can be checked, and significant departures from these assumptions can be addressed. One of the benefits of IRT, called parameter independence, relies on these assumptions holding. Parameter independence is the condition that any set of responses can be used to calibrate the item parameters, and any set of items can be used to estimate individual ability levels. This can be interpreted to mean that if students of a different year were given the CCI, the item parameters

estimated would remain unchanged. Similarly, if the same students were given a different set of test items which measured the same latent variable, conceptual knowledge of calculus, their ability estimates would remain unchanged.

To check unidimensionality, we used a method proposed by Bejar (1980) and used by Wallace and Bailey (2010). This method involves estimating difficulty parameters twice, by first dividing the test into sub-tests which may potentially measure different latent variables. On the CCI, some of the items require knowledge of derivatives, while others do not. If unidimensionality is a valid assumption, estimating the difficulty of the items requiring knowledge of derivatives should not depend on whether there are other questions on the test which do not require knowledge of derivatives. We could have chosen any set of questions which might require a different skill from those questions not chosen. We then estimate the difficulties of the items requiring knowledge of derivatives as a group, and then estimate the difficulties of those items as a part of the entire test. If unidimensionality holds, the estimates should be nearly identical. The difficulties of the nine items requiring knowledge of derivatives under each estimation method is plotted in Figure 15.

Figure 15: Item difficulty estimates in sub-test versus full test



The best fit line forms an angle of 47.9 degrees, which is quite close to the ideal 45 degrees. While Wallace and Bailey (2010) do not specify a particular statistical test at this point, they continue using a unidimensional model with angles less than 40 degrees, noting that this may indicate that the test is not actually unidimensional. Given the closeness of the angle to 45 degrees, we consider the test to be essentially unidimensional.

Local independence can be measured using Yen's (1984) Q3 statistic. This statistic is computed by comparing the predicted and model-predicted responses for each item, and then determines correlations in the residuals by item (Wallace & Bailey, 2010). The results of the Q3 statistic are shown in the table below. A value of 0.20 is suggested as a cutoff for values that should be investigated further, which are in bold.

Table 26: Yen's Q3 Statistic

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	
Q1	1.00																					
Q2	0.25	1.00																				
Q3	0.19	0.53	1.00																			
Q4	0.13	0.40	0.38	1.00																		
Q5	0.06	0.29	0.32	0.34	1.00																	
Q6	0.14	0.27	0.24	0.35	0.17	1.00																
Q7	0.07	0.22	0.26	0.24	0.22	0.14	1.00															
Q8	0.12	0.40	0.39	0.37	0.27	0.25	0.23	1.00														
Q9	0.18	0.41	0.36	0.31	0.22	0.20	0.11	0.26	1.00													
Q10	0.08	0.31	0.40	0.37	0.31	0.25	0.28	0.37	0.28	1.00												
Q11	-0.01	0.15	0.22	0.20	0.10	0.09	0.14	0.17	0.14	0.17	1.00											
Q12	0.09	0.21	0.21	0.20	0.10	0.13	0.18	0.21	0.09	0.18	0.09	1.00										
Q13	0.12	0.15	0.16	0.13	0.09	0.08	0.07	0.11	0.16	0.14	0.02	0.05	1.00									
Q14	0.07	0.23	0.23	0.29	0.14	0.16	0.17	0.22	0.17	0.18	0.10	0.12	0.09	1.00								
Q15	0.07	0.11	0.09	0.10	0.04	0.09	0.08	0.11	0.13	0.08	0.06	0.07	0.03	0.06	1.00							
Q16	0.19	0.39	0.34	0.31	0.19	0.17	0.23	0.33	0.27	0.31	0.13	0.16	0.06	0.18	0.04	1.00						
Q17	0.13	0.33	0.35	0.36	0.24	0.23	0.17	0.31	0.23	0.35	0.15	0.19	0.18	0.21	0.09	0.24	1.00					
Q18	0.06	0.18	0.11	0.14	0.03	0.06	0.04	0.15	0.13	0.12	0.03	0.10	0.12	0.08	0.04	0.08	0.11	1.00				
Q19	0.19	0.25	0.19	0.10	0.11	0.09	0.06	0.16	0.22	0.07	0.07	0.04	0.06	0.12	0.04	0.17	0.14	0.10	1.00			
Q20	0.12	0.35	0.37	0.34	0.24	0.21	0.21	0.31	0.25	0.28	0.17	0.19	0.15	0.28	0.09	0.29	0.26	0.07	0.17	1.00		
Q21	0.16	0.44	0.48	0.38	0.30	0.20	0.24	0.34	0.34	0.42	0.17	0.19	0.18	0.23	0.05	0.30	0.37	0.13	0.16	0.33	1.00	

Since almost 25% of the pairs were above the 0.20 threshold, we investigated the pairs of items to determine whether there were any traits in the pairs with high Q3 statistics which might indicate a reason for so many pairs with high Q3 values. Upon inspection, the pairs which had high Q3 values did not have any similarities, so it was decided that local independence is still a reasonable assumption in this model. Therefore, we find that the use of an IRT model is appropriate.

2.4 Computation of Gain Scores

The previous sections have considered the pretest and posttest independently. We now consider how we can determine gains achieved by students over the course of the semester. When models for the pretest and posttest are created, a score of zero is test dependent. For example, a score of zero on the pretest indicates an average ability for a student beginning college calculus. A score of zero on the posttest indicates the average ability of a student at the end of college calculus. If we were to simply estimate ability levels for each student on the pretest and the posttest, we would have a measure of gain, but not one which can be immediately understood. For example, a gain score of zero would indicate that a student's ability level, relative to the entire population, has not changed. The student's ability level has increased, only it has increased by the same amount as the entire population. A student who began slightly above average at the beginning of the semester and ended slightly below average may have achieved gains, though gained less than the average student. This student would be assigned a negative gain score under the method described above. We, instead, aimed to use a gain score which assigned positive values to those students who achieved gains, so that a gain score

of zero would indicate no gains.

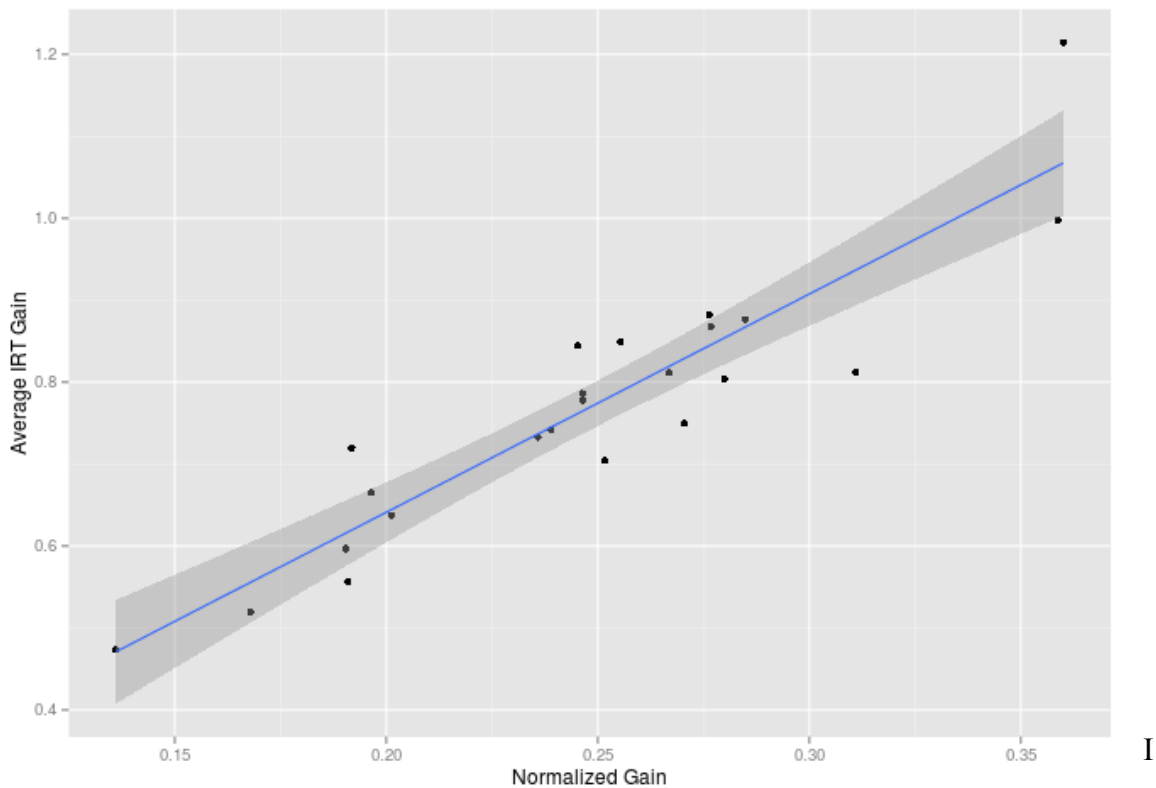
To address this concern, the ability estimates for those students taking the posttest were recomputed using the test item parameters estimated from the pretest, following the method used by Wallace and Bailey (2010). By doing this, the score of a student taking the pretest can be interpreted as being relative to the students beginning the course. For example, a student with an estimated ability of 1 on the posttest using this method would be said to have the same ability as a student beginning the course one standard deviation above the mean. By using this method, we can subtract posttest scores from pretest scores, and a positive difference indicates a gain, a negative difference indicates a loss, and a difference of 0 indicates that no change in ability occurred.

2.5 Comparison of Normalized Gain and IRT

So far we have constructed gain scores in two different ways. In Chapter 2 we used traditionally computed instructor-level normalized gain scores and in Chapter 3 we introduced the possibility of using individual-level normalized gain scores, we used normalized gain scores. Now that we have used IRT to compute gain scores in another way, we can consider how much different the normalized gain scores are from the IRT gain scores.

We initially computed normalized gain scores at the instructor-level, in the same way that normalized gain scores are usually reported. To compare, we averaged IRT gain scores by instructor, so each instructor would have two measures of gain. A comparison of these two measures of gain is displayed in Figure 16.

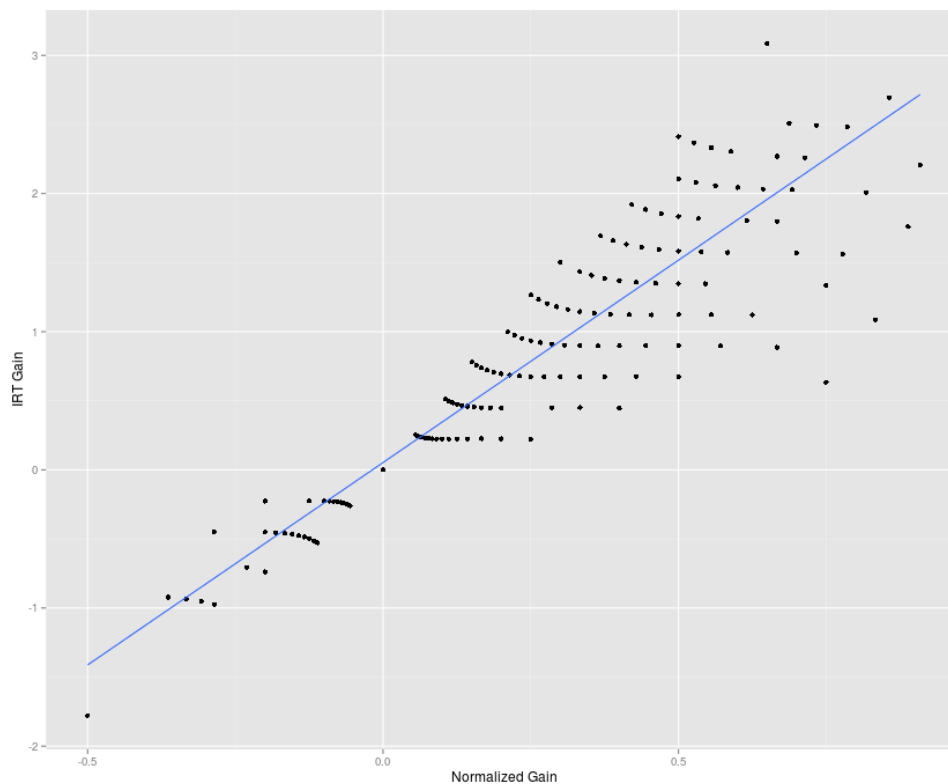
Figure 16: Normalized Gain vs. IRT Gain by Instructor



The two measures of gain at the instructor level are strongly correlated, $r(21) = 0.92$, $p < 0.01$, so 85% of the variation in one measure is explained by the other.

In Chapter 3, we introduced individual-level normalized gains, which we can directly compare to IRT gains since each are already computed at the individual-level. This relationship is displayed in Figure 17.

Figure 17: Individual Normalized Gains vs. IRT Gains



The correlation between individual normalized gains and individual IRT gains is also strong, $r(480) = 0.92$, $p < 0.01$, so $r^2 = 0.85$, suggesting that 85% of variation of one measure is accounted for by the other measure.

2.5.1 Advantages of Normalized Gain Scores

Practically, one of the fundamental differences between normalized gain scores and IRT gain scores is the way that the pretest score affects the normalized gain scores. Given two classes with the same difference between average pretest and average posttest scores, the normalized gain will be larger for the class with the higher pretest score (Wallace & Bailey, 2010). For example, consider a concept inventory with 22 questions. Suppose a class with a pretest average of 18 points achieves a posttest average of 20 points, and

another class with a pretest average of 11 points achieves a posttest average of 13 points. The normalized gain score would be higher for the first class than for the second. Even though each class improved by the same number of points, the class with the higher pretest score achieved the higher normalized gain. This dependence on pretest score can be seen directly in the formula for normalized gain,

$$\langle g \rangle = \frac{\text{Posttest score} - \text{Pretest score}}{\text{Maximum score} - \text{Pretest score}} \quad (8)$$

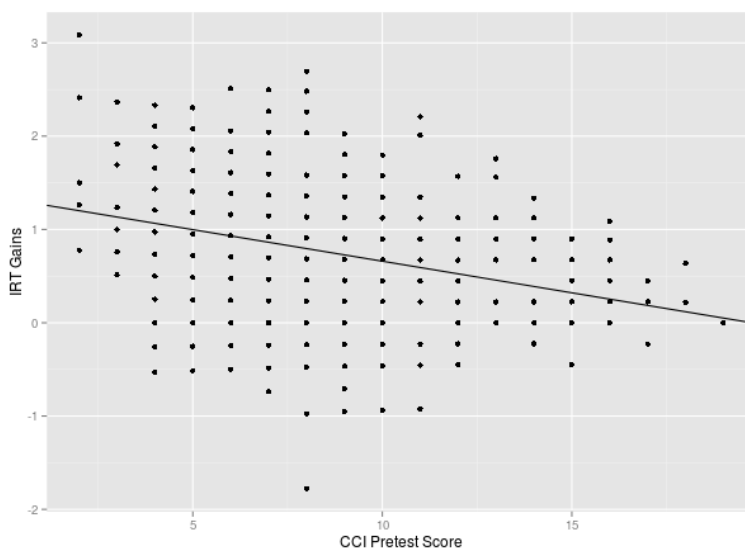
If two classes achieve the same difference between pretest and posttest on a test, the numerator will be the same. The maximum will be the same by virtue of being the same test. Having a higher pretest score will only decrease the denominator, resulting in a higher $\langle g \rangle$ value. In fact, the restriction of being the same test is not needed as the normalized gain could be equivalently defined by using percentages instead of scores and replacing the maximum score with 1.

Having a high pretest score benefit a class' gain score is a reasonable feature for a gain score to have. If the difficulties of the questions are distributed roughly normally, a class with a medium average pretest score will likely have questions very close to their ability which they were not able to successfully answer. Therefore, a small improvement in ability could result in an improvement in score. For a class with a high average pretest score, a question just beyond their ability may be further from their current ability, so an increase in score in the posttest may correspond to an increase in ability which is greater than required for the low pretest score. If the class with the high average pretest score needed to improve in ability more than the class with the low average pretest score to

achieve the same gains in points, it is reasonable for the gain score to reflect this.

The relationship between gain scores and pretest scores will be revisited in Chapter 5, but we will mention here that correlations between CCI gain scores and pretest scores follow the trends expected based on the discussion above. As discussed by Bao (2006), we expect little correlation between pretest scores and normalized gain scores because normalized gain scores factor the role of the pretest into the score. We find this to be the case, $r(480) = -0.00774$, $p = 0.8654$. When we compute the correlation between IRT gains and pretest scores, however, we find that a moderate correlation, $r(480) = -0.302$, $p < 0.001$, displayed in Figure 18.

Figure 18: IRT gains versus CCI Pretest Scores



This is important because when we are predicting gain scores in Chapter 5, the correlation between IRT gains and pretest scores indicates that CCI pretest scores may be a valuable covariate to consider in models. The independence of normalized gain scores and CCI pretest scores suggests that pretest scores would not likely be a useful predictor

of normalized gains.

2.5.2 Advantages of IRT Gains

While normalized gains have many advantages, IRT has the advantage that it is a test-independent measure of ability. This means that an instrument which consists of different questions but measures the same construct would result in the same ability estimates for individuals.

Normalized gain is the total learned out of the total that could be learned, so the full set of question on the test imposes a maximum knowledge level. Once a student has correctly answered all the questions on the instrument, the total knowledge has been achieved for that construct. IRT does not have this type of frame of reference of total knowledge, and there is no inherent maximum ability level. IRT will not be able to distinguish students if they have successfully answered all questions on the instrument, but this is a limitation of the test's ability to measure these students, and not IRT itself. Since additional questions could be asked of these students, these students could be differentiated, and no maximum ability level is ever imposed. This is largely the motivation behind the implementation of computer adaptive testing on instruments such as the GRE (Embretson & Reise, 2000). Consider the following example. Suppose version 1 of a concept inventory has 2 questions which are correctly answered by everyone who takes the test, and version 2 of the concept inventory replaces those 2 items with 2 items which are answered incorrectly by everyone on the test. In an IRT analysis, this change would not make any difference in the ability estimates. In a normalized gain analysis of the concept inventories, however, the normalized gains will be different. If the

concept inventory had 22 questions, a change from 14 to 16 ($\langle g \rangle = 0.25$) on version 1 would become a change from 12 to 14 ($\langle g \rangle = 0.2$) on version 2. This is also noteworthy when comparing normalized gains on the FCI to normalized gains on the CCI. These are two completely different instruments, and so comparing normalized gains on one with the other may not be reasonable. In particular, useful cutoffs for high-, medium-, and low- $\langle g \rangle$ scores may not transfer from one test to the other.

3 Conclusions

Both IRT gains and normalized gains aim to determine the amount of learning that has taken place during a course, but they measure this quantity in different ways. A priori, there is no objective way to choose one measure as preferred to the other, as each have advantages. IRT produces measures which are test and population independent (Embretson & Reise, 2000), and normalized gain scores are easily interpreted in terms of percent of knowledge gained. The two measures, while not perfectly related, are strongly correlated (at 0.92), suggesting that the rank ordering of students is fairly stable, regardless of the metric used. By considering both methods of computing gain scores, we are able to notice patterns that would not be found if we only utilized one method.

CHAPTER 5: ANALYSIS OF CLASSROOM INSTRUCTION USING ITEM RESPONSE THEORY

1 Relationship between IRT and IE Teaching

In Chapter 2, we introduced a protocol for measuring Interactively-Engaged instruction, which quantified teaching style without binning classes or relying on self-reporting. We constructed these measures by introducing a protocol which was used to code videos, and showed that some of the aspects of IE classrooms seemed to be correlated with gains on the CCI. In Chapter 3, we reanalyzed the IE and CCI data, recognizing the hierarchical nature of the data. Using individual gain scores (so that individual-level variables could be considered), we found that the relationship discovered in Chapter 2 appeared spurious, meaning that the relationship between IE instruction and CCI gains at the classroom level no longer existed when student-level variance was considered. Only after introducing a grouping variable which categorized students by their previous mathematics courses were the IE variables significant again.

Our results in Chapter 2 are in agreement with the result of previous studies. When our method of analysis changes, however, the results become less clear. An attempt to analyze the data using a hierarchical model (which is theoretically appropriate given the clustering in classrooms and the fact that the variable of interest, Interactive-Engagement, lies at the classroom level) put into question the initial conclusions. The latter suggests that the relationship between interactive teaching and conceptual learning may be more complicated than previously thought. When the variances associated with individual-level variables is considered, the remaining variance may then be explainable by IE variables.

One possible reason for this situation is differential benefits of IE instruction to students with different backgrounds.

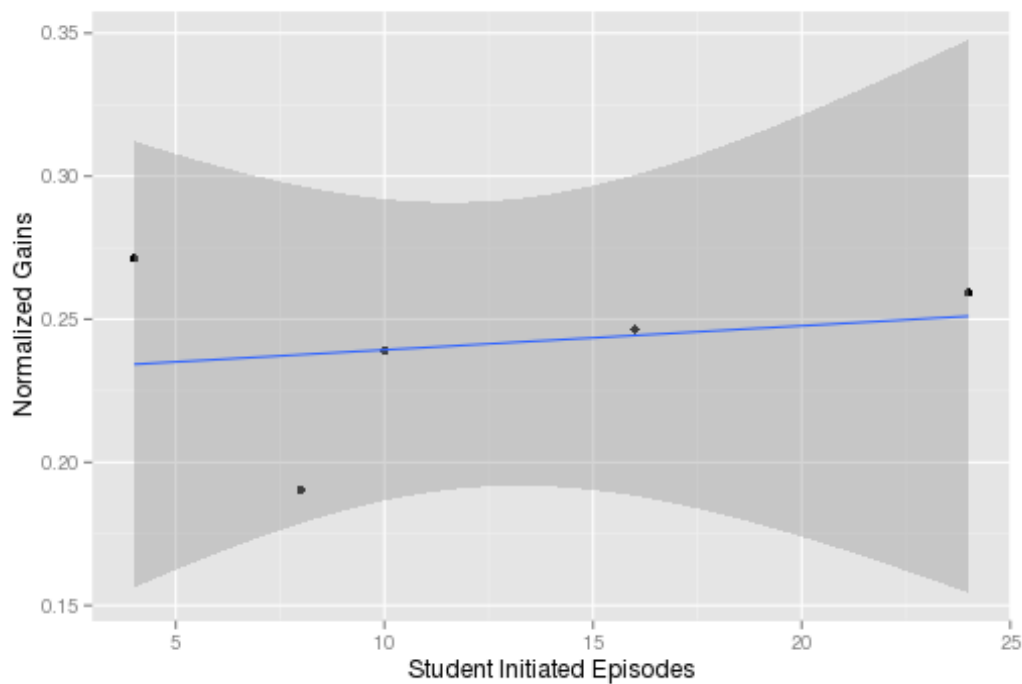
In chapter 4, we introduced an additional method for measuring gains called Item Response Theory (IRT). This was done both because IRT possesses measurement traits which are attractive, and also because we were concerned with whether the particular method of gain score measurement would affect our results. In this Chapter, we re-evaluate the effect of IE teaching styles on student learning, using the IRT gain scores constructed in Chapter 4 in addition to the normalized gain scores constructed in Chapter 2.

2 Instructor-level Results

In Chapter 2 we compared normalized gain scores at the classroom level with counts of IE episodes. Specifically, the number of revisions encouraged by the instructors and the total number of interactions were each highly correlated with student gains. To compare IRT gains with this analysis, we average the IRT gains by instructor. In this section, we revisit each of the types of counts of interactions explored in Chapter 2, considering how the use of IRT gains might change interpretations of the results in place of normalized gains.

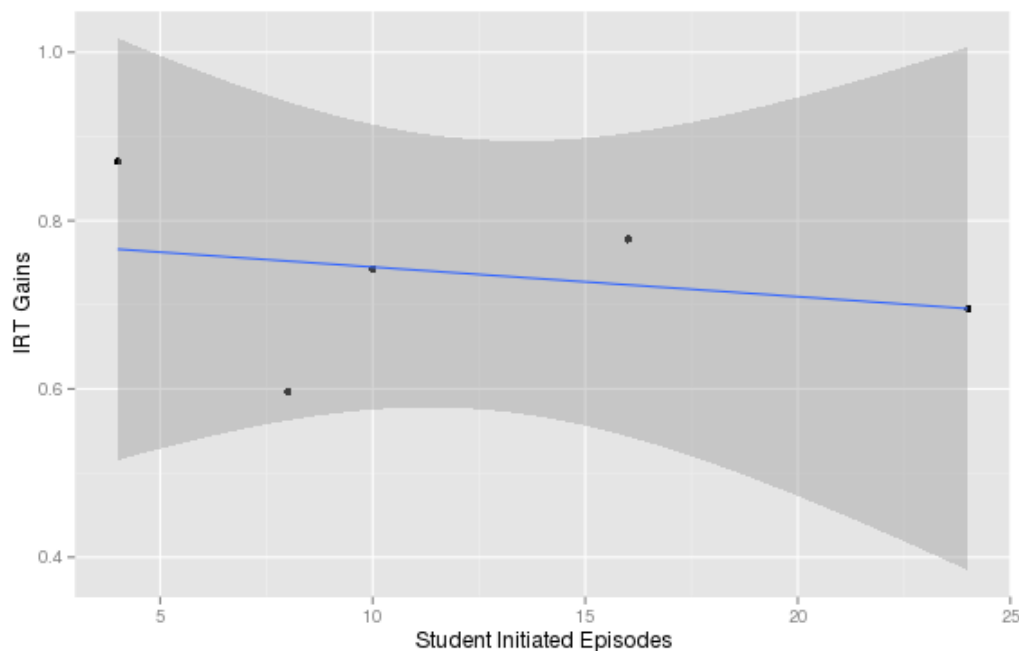
For reference, we first consider the plot of normalized gains versus the total number of student-initiated episodes. This was not a statistically significant relationship. This plot is given in Figure 19.

Figure 19: Normalized gains versus Student Initiated Episodes



IRT gains are averaged and plotted against the total number of student-initiated interaction episodes over the three classes, and the result is given in Figure 20.

Figure 20: IRT gains versus Student Initiated Episodes



The number of student-initiated episodes does not significantly predict gains using either IRT gains, $b = -0.0035$, $t(3) = -0.49$, $p = 0.66$, or by using normalized gains, $b = 0.0008$, $t(3) = 0.38$, $p = 0.73$. The dependence on pretest scores is illustrated well in this example, as gain scores are very similar between the two gain score methods with one exception. This instructor, who is represented in the scatterplot as having the most student-initiated episodes has a lower gain score when computed by IRT than when using normalized gains. This is because this class had a much higher average pretest score than the other sections, and the normalized gain score formula provides a benefit to those instructors with high pretest scores.

We next consider the total number of instructor-initiated interaction episodes. The plot of normalized gains against the number of instructor-initiated episodes is given in

Figure 21 and IRT gains against the number of instructor-initiated episodes in Figure 22.

Figure 21: Normalized gains versus Instructor Initiated Episodes

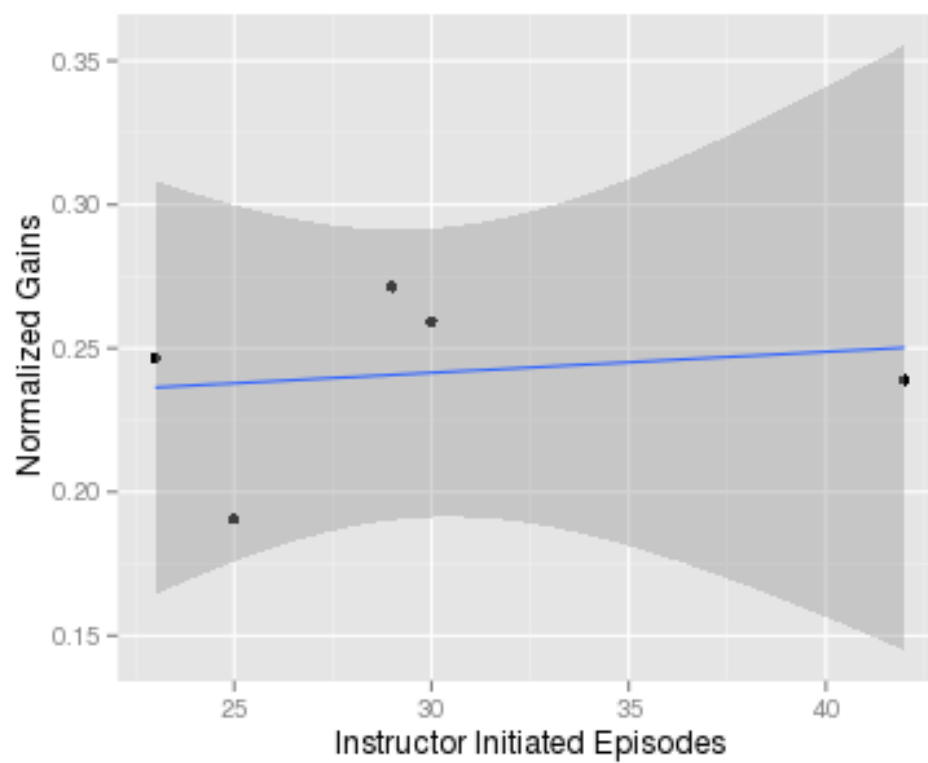
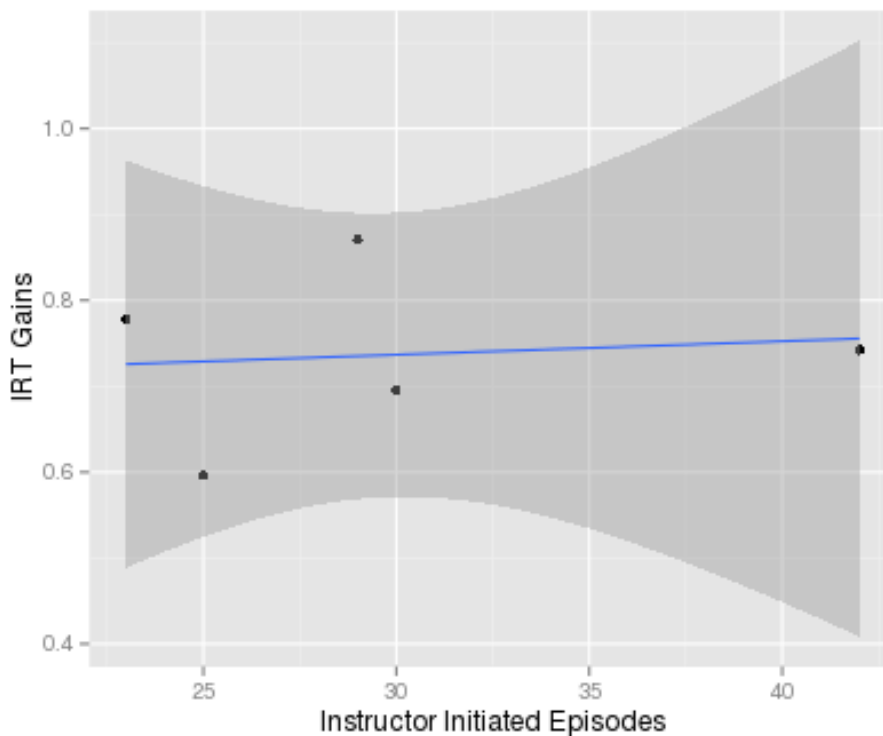


Figure 22: IRT gains versus Instructor Initiated Episodes



There is no significant relationship between instructor-initiated episodes and normalized gains, $b = 0.00073$, $t(3) = 0.31$, $p = 0.78$, or IRT gains, $b = 0.0016$, $t(3) = 0.2$, $p = 0.85$.

The next category of IE interactions considered in Chapter 2 was the total number of instructor-student interactions which took place over the three courses. The plot of normalized gains against the total number of interactions is given in Figure 23 and the plot of IRT gains against the total number of interactions is given in Figure 24.

Figure 23: Normalized gains versus all interactions

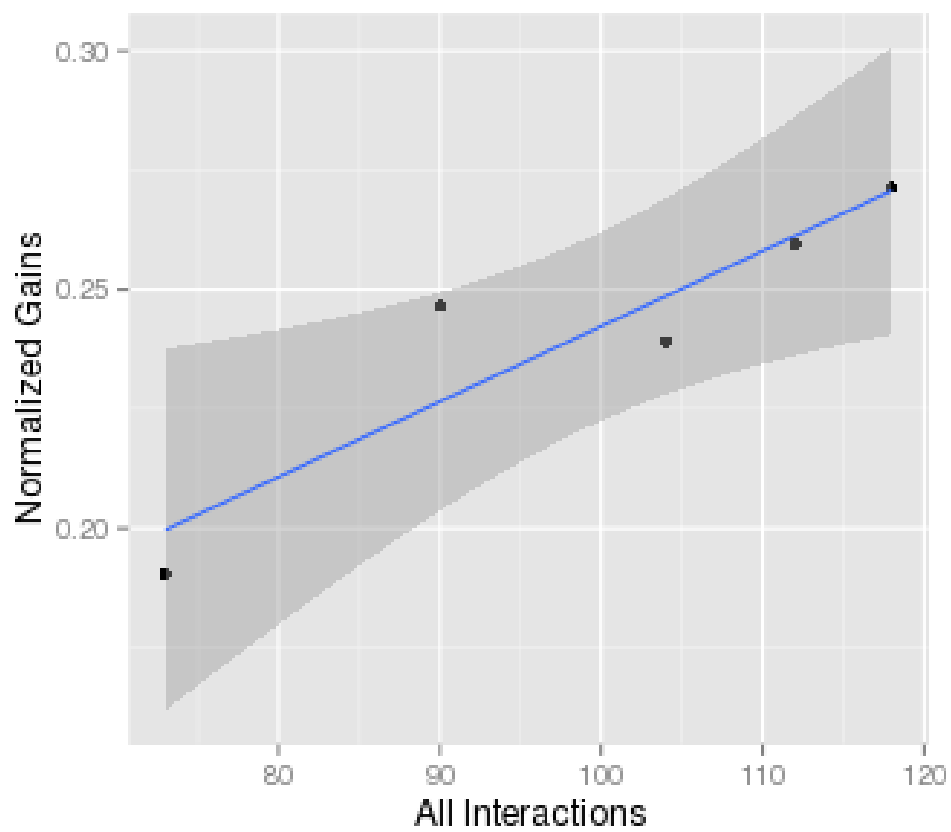
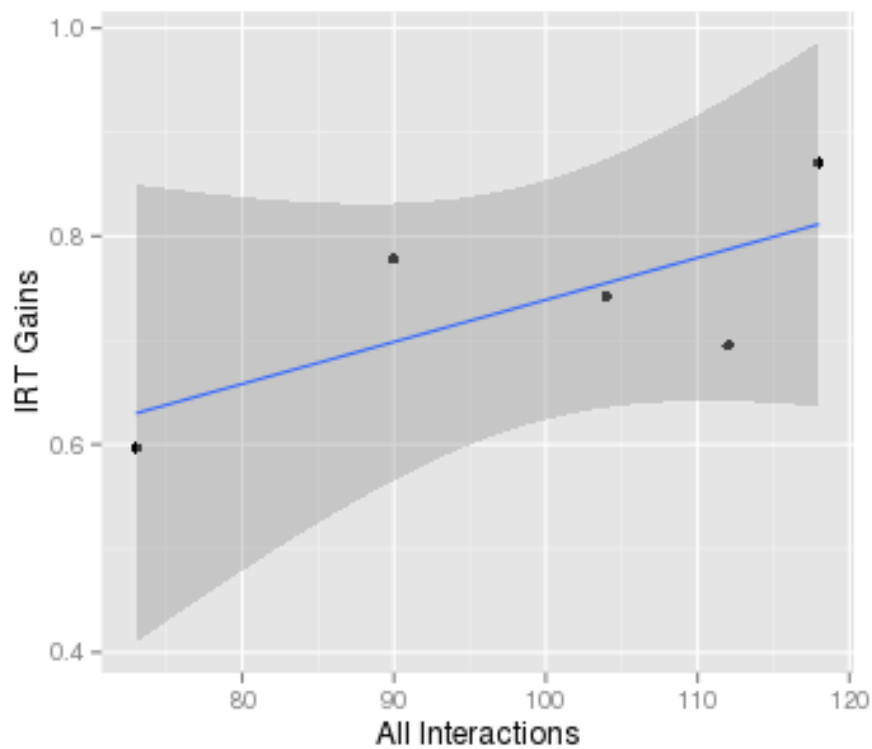


Figure 24: IRT Gains versus All Interactions



When the gains are expressed as normalized gains, the relationship is statistically significant, but when gains are measured using IRT, the gains are not significant, though the effect size is large for both metrics, seen in Table 27.

Table 27: Predicting Gains Scores Using All Interactions

Variable	B	SE(B)	β	t	Sig (p)	R ²
Model 1: All Interactions to Predict Normalized Gains						
Constant	0.0844	0.0387		2.181	0.1173	0.8491
All Interactions	0.0016	0.0004	0.9215	4.108	0.0261*	
Model 2: All Interactions to Predict IRT Gains						
Constant	0.3349	0.2249		1.489	0.233	0.5218
All Interactions	0.0040	0.0022	0.7223	1.809	0.168	

Note: B indicates the unstandardized regression coefficient. β indicates the standardized regression coefficient.

** $p < 0.05$*

Additionally, we note the standardized regression coefficients. The standardized regression coefficient when using normalized gains is nearly 1, indicating that an increase by 1 standard deviation of the number of interactions corresponds to an increase of student scores by 0.9215 standard deviations. The effect sizes for these two models as measured by Cohen's f^2 statistic are 5.627 and 1.091 respectively, both indicating a fairly large effect. This provides additional evidence that classroom-level interactions may help encourage student gains. Keeping in mind that there are only 5 instructors and so results may not generalize, this discrepancy indicates some dependence on the method of computing gain scores, namely that the statement of whether student gains on the CCI is dependent upon the number of interactions in the course, is dependent on how the gains are measured. This casts further doubt on the original relationship found in Chapter 2 between the total number of interactions and CCI gains. That relationship is only valid

provided gain scores are aggregated, and that gains are measured using normalized gains. While there is still promise in the HLM using the previous mathematics course grouping, the claim that more interactions is correlated with CCI gains seems to be an oversimplification of the relationship between the two variables.

The total number of “encouraging revisions from students” episodes is plotted against normalized gains in Figure 25 and against IRT gains in Figure 26.

Figure 25: Normalized gains versus number of Revision Encouragements

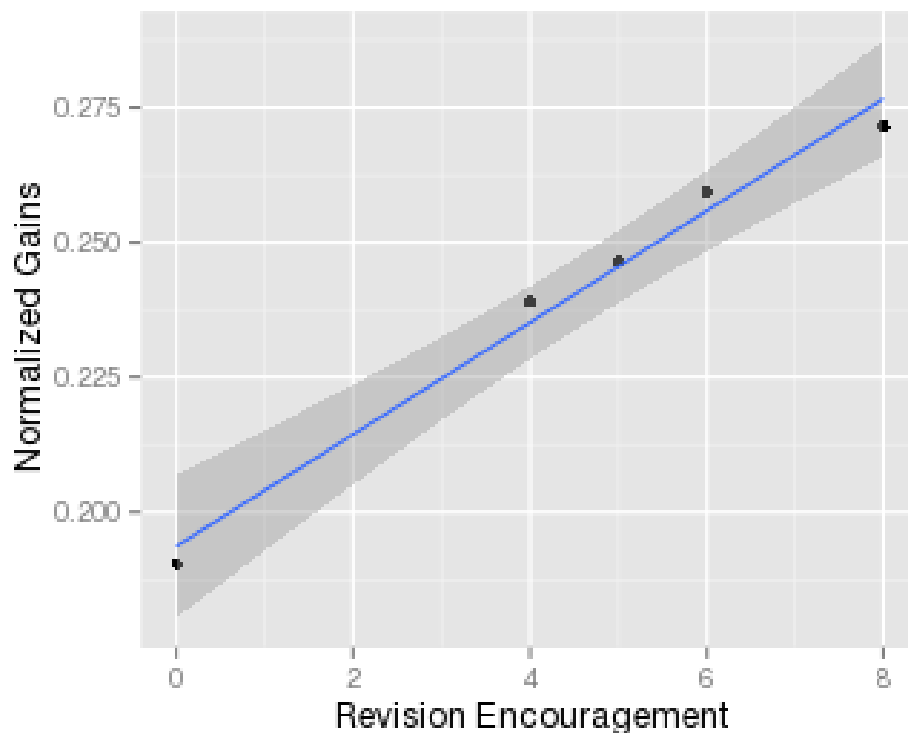
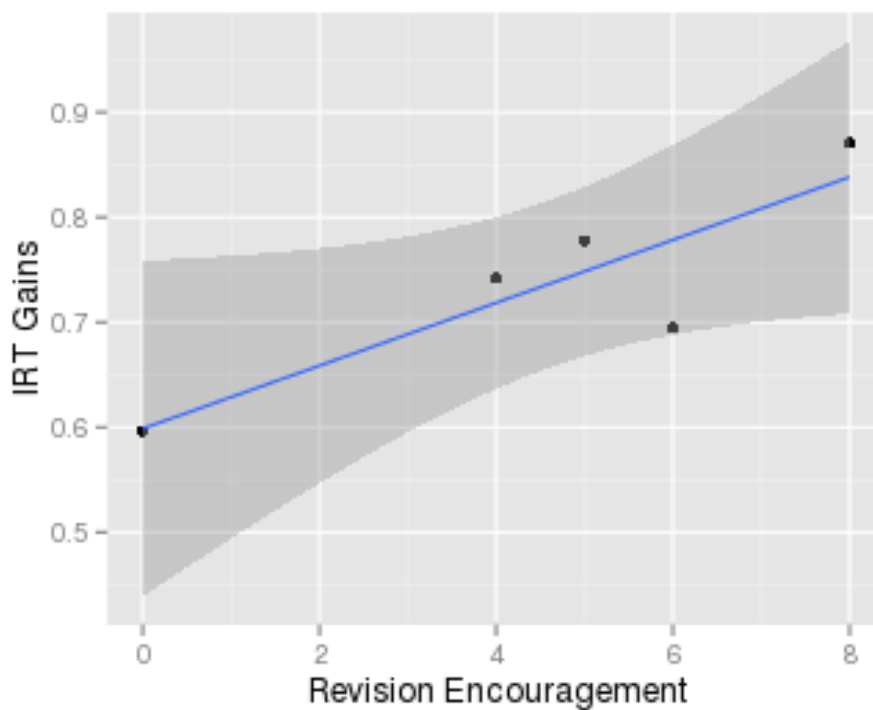


Figure 26: IRT Gains versus Number of Revision Encouragements



This measure is significant when using normalized gain scores to measure student learning, but is not so when using IRT gain scores, as seen in Table 28.

Table 28: Predicting Gains by Number of Revisions Encouraged

Variable	B	SE(B)	β	t	Sig (p)	R ²
Model 1: Revisions Encouraged to Predict Normalized Gains						
Constant	0.1936	0.0042		46.49	< 0.001*	0.9831
Number of Revisions Encouraged	0.0104	0.0008	0.9915	13.22	< 0.001*	
Model 2: Revisions Encouraged to Predict IRT Gains						
Constant	0.5987	0.0502		11.916	0.00127*	0.7695
Number of Revisions Encouraged	0.0299	0.0095	0.8772	3.164	0.05070	

Note: B indicates the unstandardized regression coefficient. β indicates the standardized regression coefficient.

* $p < 0.05$

We again note that the standardized regression coefficients indicate that changes in the number of revisions encouraged is associated with substantial changes in student scores. The effect sizes as measured by Cohen's f^2 statistic are 58.17 and 3.34, indicating that the effect of encouraging revisions on student learning seems to be very strong based on this evidence. These two values are very large for effect sizes, and so we again note that their magnitudes should be interpreted cautiously. The small sample size may be again be leading to an inflated effect size. We expect that since the IRT gains and normalized gains are highly correlated that we would not find marked differences between these two models, and this is indeed the case. The similarities between the standardized regression coefficients provides further evidence that the metric for gain scores is of great importance when determining statistical significance with this small sample size.

While encouraging revisions and the total number of interactions were the only statistically significant predictors of CCI normalized gains in Chapter 2, neither of these variables is a statistically significant predictor of IRT gains. None of the counts of other types of interactions was a significant predictor of CCI IRT gains either.

3 Hierarchical Linear Models and Individual-level Analysis using IRT Gains

We further investigate differences between normalized gains and IRT gains by considering hierarchical linear models, as was done in Chapter 3 using normalized gains. We also recall the discussion in the previous chapter that IRT gains are significantly correlated with pretest scores, so we keep CCI pretest scores included in the models. This allows us to not only consider the contributions of the pretest scores but to also consider how the other explanatory variables may correlate with pretest scores.

3.1 Null Model

In Chapter 3 it was determined that there was very little variance in normalized gain scores at the instructor-level, hence instructor-level variables alone would be unable to explain differences in student scores. We might expect that there will also be very little variance in IRT gain scores explainable at the instructor-level, despite differences in method of calculation. This is the case, as a null model predicting IRT gains, grouping students by instructor, reveals over 99.8% of the variance is at the individual level. As in Chapter 3, we consider individual-level variables which may be able to explain difference in IRT gains.

3.2 Gender

The first variable we consider as a potential predictor of gains is gender. Gender was

not a significant predictor of normalized gains, and is not a significant predictor of IRT gains either, as demonstrated in Model 1 of Table 29. If we consider gender in the same model as a measure of IE instruction, the number of revisions encouraged, we find that only the pretest score is a significant predictor of IRT gains.

Table 29: HLMs Predicting IRT Gains

<i>Fixed Effects</i>	Model 1	Model 2
Intercept (SE) t-value	1.31469 0.18553 7.086	1.2947088 0.2399631 5.395
Gender (SE) t-value	-0.10404 0.12179 -0.854	-0.3778302 0.2458071 -1.537
CCI Pretest score (SE) t-value	-0.06173 0.01905 -3.240	-0.0596310 0.0189703 -3.143
Number of revision encouragements (SE) t-value		0.0002482 0.0305351 0.008
Gender by number of revision encouragements (SE) t-value		0.0601613 0.0438966 1.371
<i>Random Effects</i>		
Intercept Variance	< 0.001	< 0.001
Residual Variance	0.44853	0.44266
% Variance Explained	< 0.001	< 0.001

These results are consistent with the results seen in using an HLM to predict normalized gain scores. We find that, as expected, instructor-level variables alone are not sufficient to explain the CCI gains observed in our data.

3.3 Individual-level Predictors

In this section, we consider the same variables as we did in Chapter 3, using IRT

gains instead of normalized gains to determine whether using IRT gains instead of normalized gains provides additional insights.

As discussed in Chapter 3, we consider previous mathematics experience as measured by previously taken mathematics courses at different level. The results of t -tests for each of the following dichotomous variables is given in the following table.

Table 30: Pairwise Comparison of IRT Gains Based on Prior Mathematics Courses

Variable	df	t -value	p -value
No Precalculus	11.514	-0.217	0.832
High School Precalculus	118.276	-1.076	0.2841
College Precalculus	177.705	2.1907	0.02978
No Calculus	296.7	-1.9392	0.05342
High School Calculus	443.738	0.4811	0.6307
College Calculus	86.473	2.1406	0.03512

We find that college precalculus and college calculus are the two significant predictors, just as was the case with normalized gains. Students who have taken college precalculus ($M = 0.635$, $SD = 0.803$) are achieving smaller gains than students who have not taken college precalculus ($M = 0.816$, $SD = 0.697$). Students who have previously taken college calculus ($M = 0.589$, $SD = 0.755$) are achieving smaller gains than those who have not ($M = 0.801$, $SD = 0.720$).

We then group students in the same way as in Chapter 3. The mean IRT gains for each group are given below.

Table 31: Mean IRT Gains of Students Based on Prior Mathematics Courses

Group	1	2	3	4
Group name	“college repeaters”	“high school repeaters”	“first-time calculus, less prepared students”	“first-time calculus, better prepared students”
Mean IRT gains	0.589	0.758	0.751	0.925

We then conduct a pairwise t-test to determine whether differences exist between the groups. The p -values of each of the comparisons is given in Table 32.

Table 32: p -values of Pairwise t -test of IRT Gains Based on Prior Mathematics Courses

Group	1	2	3
2	0.36		
3	0.42	0.94	
4	0.02	0.23	0.41

We find the same results as in Chapter 3, namely that the only difference between the four groups is between students who are repeating the course and those who are taking it for the first time.

We then compare students based on the grouping used in Chapter 3. We find the same result as in Chapter 3 when comparing whether students have taken a college-level mathematics course, $t(261.275) = 2.444$, $p = 0.015$. When comparing based on prior exposure to calculus, we find no statistically significant difference, $t(296.7) = 1.939$, $p = 0.053$. The difference between students who have not seen calculus before ($M = 0.864$, $SD = 0.803$) and those who have ($M = 0.722$, $SD = 0.681$) is more noticeable when using IRT gains than when using normalized gains, and so may be a useful area to consider in future studies.

The final models we build HLMs of IRT gains predicted by mathematics course background and IE variables along with pretest scores. In Chapter 3, we found that the IE measures of “encouraging revisions” and “all interactions” were significant predictors of normalized gains when mathematics course background was included in the model. We then use IRT gains instead of normalized gains, listed as Model 1 and Model 2 respectively in Table 33.

Table 33: HLM Predicting IRT Gains

<i>Fixed Effects</i>	Model 1	Model 2
Intercept (SE) <i>t</i> -value	0.47752 0.30449 1.568	0.057498 0.479558 0.120
Number of revisions encouraged (SE) <i>t</i> -value	0.04244 0.02185 1.942	
All Interactions (SE) <i>t</i> -value		0.006382 0.00361 1.765
Math background 2 (SE) <i>t</i> -value	0.50639 0.24774 2.044	0.505063 0.248788 2.030
Math background 3 (SE) <i>t</i> -value	0.76920 0.32484 2.368	0.763677 0.326565 2.338
Math background 4 (SE) <i>t</i> -value	0.69421 0.25590 2.713	0.685168 0.256337 2.673
Pretest score (SE) <i>t</i> -value	-0.05754 0.01883 -3.056	-0.058757 0.018889 -3.111
<i>Random Effects</i>		
Intercept Variance	< 0.001	< 0.001
Residual Variance	0.425	0.42719
% Variance Explained	< 0.001	< 0.001

Results presented in Table 33 show that neither the number of revisions encouraged nor the total number of interactions is a significant predictor of IRT gains, even when mathematics course background is included in the model. That these variables are significant when predicting normalized gain scores in an HLM but not when predicting IRT gain scores shows that the method of computing gain scores is factoring into our analysis. These differences can largely be attributed to differences in pretest scores, specifically the one classroom with higher pretest scores than the others. Without variation in pretest scores, we expect the measures to be nearly identical. With more variation in pretest scores, we expect there to be greater differences between the two gain score measures, affecting the relationship with IE indicators. The negative coefficient of pretest score when predicting IRT gains is not surprising. While IRT does not impose a limit on an individual's ability level, a particular test does. One application of IRT, computer adaptive testing, uses this fact to great advantage by picking the questions to be the most appropriate for the person taking the test. On a paper test, however, an individual's maximum ability level is determined as the score achieved when all questions are answered correctly. The higher one's ability, then, the less one is able to gain when using IRT to measure ability level. As pretest scores increase, the less one would be expected to achieve in gains.

4 Conclusions

In attempting to predict student gains, we have discussed two different methods for measuring gains, namely normalized gains and IRT gains. When comparing at the

classroom level, as is typically done in studies involving concept inventories, we found that the total number of revisions encouraged and the total number of interactions each significantly predicted normalized gains. When a hierarchical linear model was created to account for individual-level variance, these relationships were no longer significant. When prior mathematics courses were considered in the model, the connection between these measures of IE and normalized gains was again significant. IRT was then introduced as a different method for computing gains. When IRT gains are predicted using the same IE variables as were used to predicted normalized gains, all relationships are non-significant.

CHAPTER 6: CONCLUSIONS AND IMPLICATIONS

1 General Comments

When Hake (1998) conducted his study, he suggested that IE teaching may be more effective for some students than others. He hinted that it would be useful to investigate individual student characteristics:

There is commonly a large spread in g 's for individual students in a course, with g 's ranging from the maximum $g = 1.0$ to $g = 0.0$ (or even negative). Why are current IE methods relatively effective for some student and ineffective for others? To help answer these questions it would be useful to carry out, for any given course, in-depth studies of students in the lower- $g < 0.3$ and higher- $g > 0.6$ ranges: e.g., (a) GPA's and SAT's, (b) educational backgrounds, (c) evaluations by teachers, (d) interviews by physics-education researchers, (e) study habits, (f) views on science and learning, (g) attitudes towards the course, and (h) math skills. (p. 26)

While he did consider individual-level variables to be important, they were never included in his analysis. This study addresses that possibility, while further investigating the ways that IE instruction and CCI gains can be studied. Specifically, it provides a framework for analyzing the effect that IE teaching has on specific groups of students, as the gender and prior mathematics class analysis did by using hierarchical linear modeling.

The results of this study suggest that the connection between IE instruction and CCI gains may be more complicated than has been previously asserted. These implications are considered in two sections below, as implications for future research and implications for teaching.

2 Implications for Future Research

Research into concept inventories, especially those that focus on IE instruction, have primarily used the same methods of analysis. This typically includes using instructor-

level normalized gain scores and either grouping instructors as either using IE or TL instructional techniques or grouping by having instructors self-identify the percentage of time spent in IE activities. As has been discussed earlier, reliance on self-identification introduces possible bias, and so the protocol developed in this study contributes a more objective way to measure this particular type of instructional technique.

HLM allows for a more in-depth analysis of our data by allowing us to consider individual-level variables, and we recommend as a technique to use in future studies of CCI gains. That an HLM indicated the relationship between IE instruction and CCI gains which was apparent in Chapter 2 may have been spurious provides further evidence that simply analyzing gains at the classroom level may miss important conclusions.

The results of our IRT analysis provided an additional method for computing gains which has a different set of advantages from normalized gains, and also provided information about the instrument itself. Despite Epstein's (2007) validation of the CCI, we found that one particular item was not following expected patterns. While we were not able to find a reason for this behavior, there may be some reason the specific student population in our study was misinterpreting the question. Student interviews may resolve this issue, but we would not even be aware of the issue if we only computed normalized gain scores. For this reason, we believe that the analysis techniques used in this study provide advantages over those typically used, and suggest they be used in future studies.

3 Implications for Teaching

The results of the study suggest that the relationship between IE instruction and CCI gains are more complicated than previously thought. The claim that the level of IE

instruction correlates with CCI gains is dependent upon the method of analysis, and so cannot be definitively answered based on the results of this study. However, the measures of effect size found in this study do suggest that, despite the small sample size, there does seem to be evidence that at least some of the measures of IE predict student gains.

One of the most interesting findings of this study is the significance of measures of IE when previous mathematics backgrounds are controlled for. As previously indicated, which instructor a student has is not correlated with the students' mathematics course background, and since IE variables are defined at the classroom level, this is not simply instructors teaching different student populations. There are two reasons we might find results such as these. First, IE instruction, and interactive instruction in general, can largely be viewed in a constructivist framework. The purpose of these instructional techniques is to encourage students to develop their own knowledge. If an instructor intends to successfully encourage students to develop knowledge, one beneficial tool would be knowing students' background and current level of understanding. By acknowledging students' backgrounds, an instructor may be able to tailor instruction to their specific students. It would, then, not be surprising that prior mathematics courses would be a significant variable to include in a model. Second, it may be the case that IE instruction is more effective for some student populations based on their prior mathematics courses. It may be the case that the backgrounds of the students in the class will determine how effective IE instruction will be for that particular group of students. To determine whether this is the case, however, we would need a study which included instructors teaching using varying levels of IE and having different student makeups. A

larger scale study with multiple instructors and multiple classes for each instructor would allow these effects to be studied as a single instructor's instructional styles could be studied in different settings with their effects on different groups of students.

4 Future Directions

There are three directions that this study could be extended to better understand the phenomenon of IE instruction.

1. It may take a longer amount of time than one semester to see the effects of IE instruction, and the potential effect it has on student understanding. By conducting a longitudinal study over multiple years, it may be possible to find additional effects of IE instruction. These effects might take the form of greater conceptual understanding which occurs later, or is retained for a longer time. Additionally, other variables such as attitudes towards mathematics, conceptual understanding in other sciences, or persistence in STEM fields could be studied.
2. The categories developed in the coding protocol were based on definitions of IE present in the literature, and the development of these codes to determine what episodes should count was developed by two researchers. Some notions, such as sensemaking, may be conceptualized differently by instructors than by students. Student interviews may provide additional insights as to the traits of interactions which they find help them make sense of the material. This could be particularly useful as it may illustrate ways that instructors and students think differently about making sense of content, and may help instructors cater their sensemaking talk in a way that is most useful to students.

3. Studies in physics of IE instruction and closely related instructional styles have indicated that conceptual knowledge can be improved without any loss of procedural skill (Mazur, 1997). By studying students' responses on traditional assessments which focus on procedural skill, insights as to whether this is also true in mathematics.

REFERENCES

- Adelman, C. (2006). *The Toolbox Revisited: Paths to Degree Completion from High School Through*. Department of Education. Citeseer.
- Albert, D., & Lukas, J. (1999). *Knowledge spaces: theories, empirical research, and applications*. Mahwah, NJ [u.a.]: L. Erlbaum.
- Allen, K. (2006). *The Statistics Concept Inventory: Development and analysis of a cognitive assessment instrument in statistics*. (Unpublished doctoral dissertation). University of Oklahoma.
- Allen, K. (2007). Getting More from Your Data: Application of Item Response Theory to the Statistics Concept Inventory. *2007 ASEE Annual Conference and Exposition*.
- Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, 39(10), 952–978. doi:10.1002/tea.10053
- Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical care*, 42(1, Supplement), 7–16. doi:10.1097/01.mlr.0000103528.48582.7c
- Aslanides, J., & Savage, C. (2013). The Relativity Concept Inventory: development, analysis and results. *arXiv preprint arXiv:1302.7094*, 1–19.
- Ball, D. L., Ferrini-Mundy, J., Kilpatrick, J., Milgram, R. J., Schmid, W., & Schaar, R. (2005). Reaching for common ground in K-12 mathematics education. *Notices of the AMS*, 52(9), 1055–1058.
- Bao, L. (2006). Theoretical comparisons of average normalized gain calculations. *American Journal of Physics*, 74(10), 917. doi:10.1119/1.2213632
- Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using Eigen and Eigen and S4 classes.
- Bejar, I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement*, 17(4), 283–296.
- Berrett, D. (2012a, February 19). How “Flipping” the Classroom Can Improve the Traditional Lecture. *The Chronicle of Higher Education*.
- Berrett, D. (2012b, October 25). Lectures Still Dominate Science and Math Teaching,

Sometimes Hampering Student Success. *The Chronicle of Higher Education*.

- Bonsangue, M., & Drew, D. (1995). Increasing minority students' success in calculus. *New Directions for Teaching*, 61, 23–33.
- Bressoud, D. M. (2011). The worst way to teach. *MAA Launchings*, July.
- Buck, J. (2005). Active and cooperative learning in signal processing courses. *Signal Processing Magazine, IEEE*, 22(March), 76–81.
- Carey, B. B. (2011, May 12). Less Talk, More Action: Improving Science Learning. *The New York Times*.
- Carlson, M., Madison, B., & West, R. (2010). The Calculus Concept Readiness (CCR) Instrument: Assessing student readiness for calculus. History and Overview.
- Carlson, M., Oehrtman, M., & Engelke, N. (2010). The Precalculus Concept Assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction*, 28(2), 113–145. doi:10.1080/07370001003676587
- Carmichael, S. B., Martino, G., Porter-Magee, K., & Wilson, W. S. (2010). *The State of State Standards - and the Common Core - in 2010*.
- Clark, M., & Lovric, M. (2009). Understanding secondary-tertiary transition in mathematics. *International Journal of Mathematical Education in Science and Technology*, 40(6), 755–776. doi:10.1080/00207390902912878
- Cline, K., & Lomen, D. (2009). Classroom voting: Active learning in differential equations. *Consortium of Ordinary Differential Equations Journal*, 1–6.
- Cobb, P., Wood, T., & Yackel, E. (1990). Chapter 9: Classrooms as learning environments for teachers and researchers. *Journal for research in Mathematics Education Monograph*, 4(1990), 125–210.
- Code, W., Kohler, D., Piccolo, C., & Maclean, M. (2012). Teaching Methods Comparison in a Large Introductory Calculus Class. *15th Annual Conference on Research in Undergraduate Mathematics Education* (pp. 375–379). Portland, OR.
- Cohen, E. G. (1994). *Designing groupwork: Strategies for the heterogeneous classroom* (Second.). New York, NY: Teachers College Press.
- Coletta, V. P., & Phillips, J. a. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics*, 73(12), 1172. doi:10.1119/1.2117109

- Cooney, T. (1985). A beginning teacher's view of problem solving. *Journal for Research in Mathematics Education*, 16(5), 324–336.
- Crouch, C. H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics*, 69(9), 970. doi:10.1119/1.1374249
- Crouch, C. H., Watkins, J., & Fagen, A. (2007). Peer instruction: engaging students one-on-one, all at once. *Research-Based Reform of*, 1–55.
- Deshler, J. M. (2009). *Predicting student performance in core math classes*. (Unpublished doctoral dissertation). The University of New Mexico.
- Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science (New York, N.Y.)*, 332(6031), 862–864. doi:10.1126/science.1201783
- Doignon, J.-P., & Falzmagne, J.-C. (1999). *Knowledge spaces*. Berlin: Springer.
- Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, N.J: L. Erlbaum Associates.
- Epstein, J. (2007). Development and validation of the Calculus Concept Inventory. *Proceedings of the Ninth International Conference on Mathematics Education in a Global Community* (pp. 165–170).
- Fagen, A., & Crouch, C. H. (2002). Peer instruction: Results from a range of classrooms. *Physics Teacher*, 40(April), 206–209.
- Falzmagne, J.-C., Koppen, M., Villano, M., Doignon, J.-P., & Johannesen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review*, 97(2), 201–224.
- Favia, A., Comins, N. F., & Thorpe, G. L. (2012). The Elements of Item Response Theory and its Framework in Analyzing Introductory Astronomy College Student Misconceptions. I. Galaxies, 43. *Physics Education*.
- Fennema, E., & Franke, M. L. (1992). Teachers' knowledge and its impact. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 147–164). New York: Macmillian.
- Franke, M. L., Kazemi, E., & Battey, D. (2007). Mathematics teaching and classroom practice. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 225–256). Charlotte, NC: Information Age Publishing.

- Froyd, J. E. (2007). Evidence for the efficacy of student-active learning pedagogies. *Project Kaleidoscope*, 66(1), 64–74.
- Fullilove, R. E., & Treisman, P. U. (1990). Mathematics achievement among African American undergraduates at the University of California, Berkeley: An evaluation of the mathematics workshop program. *Journal of Negro Education*, 59(3), 463–478.
- Garvin-Doxas, K., Klymkowsky, M., & Elrod, S. (2007). Building, using, and maximizing the impact of concept inventories in the biological sciences: Report on a National Science Foundation sponsored conference on the construction of concept inventories in the biological sciences. *CBE—Life Sciences Education*, 6(4), 277. doi:10.1187/cbe.07
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- Hagedorn, L. S., Siadat, M. V., Fogel, S. F., Nora, A., & Pascarella, E. T. (1999). Success in college mathematics: Comparisons between remedial and nonremedial first-year college students. *Research in Higher Education*, 40(3), 261–284.
- Hagerty, G., & Smith, S. (2005). Using the Web-Based Interactive Software ALEKS to Enhance College Algebra. *Mathematics and Computer Education*, 39(3), 12.
- Hake, R. R. (1998a). 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. doi:10.1119/1.18809
- Hake, R. R. (1998b). Interactive-engagement methods in introductory mechanics courses. *Physics Education Research*, 74, 64–74.
- Hake, R. R. (2007). Six lessons from the physics education reform effort. *Latin American Journal of Physics Education*, 1(1), 24–31.
- Halloun, I. (1985). The initial knowledge state of college physics students. *American Journal of Physics*, 53(11), 1043. doi:10.1119/1.14030
- Halloun, I., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics*, 53(11), 1056. doi:10.1119/1.14031
- Hazari, Z., Tai, R. H., & Sadler, P. M. (2007). Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors. *Science Education*, 91(6), 847–876. doi:10.1002/sc.20223
- Heller, P., & Huffman, D. (1995). Interpreting the force concept inventory: A reply to

- Hestenes and Halloun. *The Physics Teacher*, 33, 507–511.
- Hestenes, D., & Halloun, I. (1995). Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller. *Physics Teacher*, 33(8), 502–504.
- Hestenes, D., & Wells, M. (1992). A mechanics baseline test. *The Physics Teacher*, 30, 159–166. doi:10.1119/1.2343498
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30(3), 141–158. doi:10.1119/1.2343497
- Hiebert, J. (2003). What research says about the NCTM standards. *A research companion to principles and standards for school mathematics* (pp. 5–23). Reston, VA: National Council of Teachers of Mathematics.
- Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. 1, pp. 371–404). Charlotte, NC: Information Age.
- Hmelo-Silver, C. E., Golan Duncan, R., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, 42(2), 99–107.
- Hufferd-Ackles, K., Fuson, K. C., & Sherin, M. G. (2004). Describing levels and components of a math-talk learning community. *Journal for Research in Mathematics Education*, 35(2), 81–116. doi:10.2307/30034933
- Huffman, D., & Heller, P. (1995). What Does the Force Concept Inventory Actually Measure? *Physics Teacher*, 33(3), 138–43.
- Hughes Hallett, D. (2006). What have we learned from calculus reform? The road to conceptual understanding. *MAA NOTES*, 69(July), 43.
- Hughes Hallett, D., Robinson, M., & Lomen, D. (2005). *Conceptests: Active learning in calculus. Mathematics Education into the 21st Century Project*.
- Kersting, N. B., Givvin, K. B., Sotelo, F. L., & Stigler, J. W. (2009). Teachers' Analyses of Classroom Video Predict Student Learning of Mathematics: Further Explorations of a Novel Measure of Teacher Knowledge. *Journal of Teacher Education*, 61(1-2), 172–181. doi:10.1177/0022487109347875
- Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring Usable Knowledge: Teachers' Analyses of Mathematics Classroom

Videos Predict Teaching Quality and Student Learning. *American Educational Research Journal*, 49(3), 568–589. doi:10.3102/0002831212437853

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75–86.

Koirala, H. P. (1997). Teaching of calculus for students' conceptual understanding. *The Mathematics Educator*, 2(1), 52–62.

Kost, L., Pollock, S. J., & Finkelstein, N. (2009). Characterizing the gender gap in introductory physics. *Physical Review Special Topics - Physics Education Research*, 5(1), 010101. doi:10.1103/PhysRevSTPER.5.010101

Laursen, S., Hassi, M. L., Kogan, M., Hunter, A. B., & Weston, T. (2011). *Evaluation of the IBL Mathematics Project: Student and Instructor Outcomes of Inquiry-Based Learning in College Mathematics*. Learning. Boulder, CO.

Libarkin, J. (2008). Concept inventories in higher education science. *National Research Council Promising Practices in Undergraduate STEM Education Workshop 2*. Washington, D. C.

Long, M. C., Iatarola, P., & Conger, D. (2009). Explaining Gaps in Readiness for College-Level Math: The Role of High School Courses. *Education Finance and Policy*, 4(1), 1–33. doi:10.1162/edfp.2009.4.1.1

Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, 74(2), 118. doi:10.1119/1.2162549

Malone, K. (2008). Correlations among knowledge structures, force concept inventory, and problem-solving behaviors. *Physical Review Special Topics - Physics Education Research*, 4(2), 1–15. doi:10.1103/PhysRevSTPER.4.020107

Mann, W. R. (1976). Some disquieting effects of calculus in high school. *The High School Journal*, 59(6), 237–239.

Marbach-Ad, G., Briken, V., El-Sayed, N. M., Frauwirth, K., Fredericksen, B., Hutcheson, S., Gao, L.-Y., et al. (2009). Assessing student understanding of host pathogen interactions using a concept inventory. *Journal of Microbiology & Biology Education*, 10(1), 43–50. doi:10.1128/jmbe.v10.98

Marbach-Ad, G., McAdams, K. C., Benson, S., Briken, V., Cathcart, L., Chase, M., El-Sayed, N. M., et al. (2010). A model for using a concept inventory as a tool for

- students' assessment and faculty professional development. *Life Sciences Education*, 9, 408–416. doi:10.1187/cbe.10
- Mazur, E. (1997). *Peer instruction: A user's manual*. Upper Saddle River, NJ: Prentice Hall.
- Mazur, E., & Watkins, J. (2010). Just-in-Time Teaching and Peer Instruction. In S. P. Simkins & M. H. Maier (Eds.), *Just-in-Time Teaching: Across the Disciplines, Across the Academy* (pp. 39–62). Sterling, VA: Stylus Publishing, LLC.
- Meltzer, D. E. (2002). The relationship between mathematics preparation and conceptual learning gains in physics: A possible “hidden variable” in diagnostic pretest scores. *American Journal of Physics*, 70(12), 1259. doi:10.1119/1.1514215
- Michael, J. (2006). Where's the evidence that active learning works? *Advances in Physiology Education*, 30, 159–167. doi:10.1152/advan.00053.2006
- Miller, R. L., Santana-Vega, E., & Terrell, M. (2006). Can good questions and peer discussion improve calculus instruction? *Primus*, 16(3), 193–203. doi:10.1080/10511970608984146
- Mislevy, B. (n.d.). Mislevy Hake Emails. Retrieved from <http://www.education.umd.edu/EDMS/mislevy/papers/Gain/>
- Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. a. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, 330(6008), 1234–7. doi:10.1126/science.1195996
- Monson, D. S. (2011). *The Relationship Between Beliefs and Practices of Mathematics Teachers Who Use a Standards-Based Curriculum*. UNIVERSITY OF MINNESOTA.
- Moreno, S. E., & Muller, C. (1999). Success and diversity: The transition through first-year calculus in the university. *American Journal of Education*, 108(1), 30. doi:10.1086/444231
- Morris, G. A., Harshman, N., Branum-Martin, L., Mazur, E., Mzoughi, T., & Baker, S. D. (2012). An item response curves analysis of the Force Concept Inventory. *American Journal of Physics*, 80(9), 825–831. doi:10.1119/1.4731618
- Mulford, D. R., & Robinson, W. R. (2002). An inventory for alternate conceptions among first-semester general chemistry students. *Journal of Chemical Education*, 79(6), 739. doi:10.1021/ed079p739

- Norton, A., McCloskey, A., & Hudson, R. a. (2011). Prediction assessments: Using video-based predictions to assess prospective teachers' knowledge of students' mathematical thinking. *Journal of Mathematics Teacher Education*, *14*(4), 305–325. doi:10.1007/s10857-011-9181-0
- Novak, G. M., Gavrin, A. D., Christian, W., & Patterson, E. T. (1999). *Just in time teaching*.
- Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257. doi:10.3102/01623737026003237
- Panel, C. (1987). Report of the CUPM Panel on Calculus Articulation: Problems in Transition from High School Calculus to College Calculus. *The American Mathematical Monthly*, *94*(8), 776–785.
- Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist*, *36*(2), 89–101. doi:10.1207/S15326985EP3602
- Pilzer, S. (2001). Peer instruction in physics and mathematics. *Primus*, *11*(2), 185–192. doi:10.1080/10511970108965987
- Pollock, S. J., & Finkelstein, N. (2007). Reducing the gender gap in the physics classroom: How sufficient is interactive engagement? *Physical Review Special Topics-Physics*, (March), 1–4. doi:10.1103/PhysRevSTPER.3.010107
- Prather, E., & Brissenden, G. (2008). Development and application of a situated apprenticeship approach to professional development of astronomy instructors. *Astronomy Education Review*, *7*(2), 1. doi:10.3847/AER2008016
- Prather, E., Rudolph, A. L., & Brissenden, G. (2009). Teaching and learning astronomy in the 21st century. *Physics Today*, *62*(10), 41. doi:10.1063/1.3248478
- Prather, E., Rudolph, A. L., Brissenden, G., & Schlingman, W. M. (2009). A national study assessing the teaching and learning of introductory astronomy. Part I. The effect of interactive instruction. *American Journal of Physics*, *77*(4), 320. doi:10.1119/1.3065023
- Raudenbush, S. (1993). Hierarchical linear models and experimental design. In L. K. Edwards (Ed.), *Statistics: Textbooks and monographs, Vol. 137* (pp. 459–496). New York, NY: Marcel Dekker.
- Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: Sage Publications.

- Raymond, A. M. (1997). Inconsistencies between a Beginning Elementary School Teacher's Mathematics Beliefs and Teaching Practice. *Journal for Research in Mathematics Education*, 28(5), 550–576.
- R Core Team. (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
- Redish, E. F. (1994). Implications of cognitive studies for teaching physics. *American Journal of Physics*.
- Redish, E. F., & Steinberg, R. N. (1999). Teaching physics: Figuring out what works. *Physics Today*, 52, 24–31.
- Reynolds, N., & Conaway, B. (2003). Factors affecting mathematically talented females' enrollment in high school calculus. *Journal of Secondary Gifted Education*, 14(4), 218–228.
- Rhea, K. (n.d.). The Calculus Concept Inventory at a large research university. *Unpublished manuscript*.
- Rhoads, T. R., & Roedel, R. J. (1999). The Wave Concept Inventory - A cognitive instrument based on Bloom's Taxonomy. *Proceedings of the 29th ASEE/IEEE Frontiers in Education Conference* (Vol. 3, pp. 14–18). IEEE.
- Riegler, P. (2010). Towards mathematics education research - does physics education research serve as a model? *sefi.htw-aalen.de*.
- Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of educational psychology*, 91, 175–189.
- Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, 93(2), 346–362.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25.
- Roach, K., Roberson, L., Tsay, J. J., & Hauk, S. (n.d.). Mathematics graduate teaching assistants' question strategies. *Proceedings of the 13th Annual Conference on Research in Undergraduate Mathematics Education*.
- Roberts, J., & Monaco, J. P. (2006). Effect size measures for the two-level linear multilevel model. *Annual meeting of the American Educational Research*

Association.

- Rudolph, A. L., Prather, E., Brissenden, G., Consiglio, D., & Gonzaga, V. (2010). A national study assessing the teaching and learning of introductory astronomy part II: The connection between student demographics and learning. *Astronomy Education Review*, 9(3), 010107.
- Savinainen, A., & Scott, P. (2002). The Force Concept Inventory: a tool for monitoring student learning. *Physics Education*, 37(1), 45–52. doi:10.1088/0031-9120/37/1/306
- Schlingman, W., Prather, E., & Wallace, C. S. (2012). A classical test theory analysis of the Light and Spectroscopy Concept Inventory national study data set. *Astronomy Education*.
- Schoenfeld, A. H. (1995). A brief biography of calculus reform. *UME Trends: News and Reports on Undergraduate Mathematics Education*, 6(6), 3–5.
- Schoenfeld, A. H. (2004). The math wars. *Educational Policy*, 18(1), 253–286. doi:10.1177/0895904803260042
- Selvin, P. (1992). Math education: Multiplying the meager numbers. *Science*, 258(5085), 1200–1201.
- Seymour, E., & Hewitt, N. M. (1997). *Talking about leaving: Why undergraduates leave the sciences*. Boulder, Colo: Westview Press.
- Sfard, A., Neshet, P., Streefland, L., Cobb, P., & Mason, J. (1998). Learning mathematics through conversation: Is it as good as they say? *For the Learning of Mathematics*, 18(1), 41–51.
- Smith, A. C., Stewart, R., Shields, P., Hayes-Klosteridis, J., Robinson, P., & Yuan, R. T. (2005). Introductory biology courses: A framework to support active learning in large enrollment introductory science courses. *Cell Biology Education*, 4(2), 143–56. doi:10.1187/cbe.04-08-0048
- Stein, M. K., Engle, R., Smith, M., & Hughes, E. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning*, 10(4), 313–340. doi:10.1080/10986060802229675
- Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, 2(1), 50–80. doi:10.1080/1380361960020103

- St. Jarre, K. (2008). They Knew Calculus when They Left: The Thinking Disconnect between High School and University. *Phi Delta Kappan*, 90(2), 4.
- Tartre, L. A., & Fennema, E. (1995). Mathematics achievement and gender: A longitudinal study of selected cognitive and affective variables [Grades 6–12]. *Educational Studies in Mathematics*, 28(3), 199–217.
- Terrell, M. (2003). Asking good questions in the mathematics classroom. *Mathematicians and Education Reform Forum Newsletter* (Vol. 15, pp. 3–5).
- Thompson, A. G., Philipp, R. A., Thompson, T. W., & Boyd, B. A. (1994). Computational and conceptual orientations in teaching mathematics. In A. Coxford (Ed.), *1994 Yearbook of the NCTM* (pp. 79–92). Reston, VA: NCTM.
- Treisman, U. (1992). Studying students studying calculus: A look at the lives of minority mathematics students in college. *College Mathematics Journal*, 23(5), 362–372.
- Wallace, C. S., & Bailey, J. M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review*, 9(1). doi:10.3847/AER2010024
- Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics*, 78(10), 1064. doi:10.1119/1.3443565
- Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Media. New York, NY: Springer.
- Wieman, C. (2007). Why not try a scientific approach to science education? *Change: The Magazine of Higher Learning*, 39(5), 9–15.
- Wu, H. (1999). Basic skills versus conceptual understanding. *American Educator*, 23(3), 14–19.
- Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125–145.