

Examining issues of underrepresented minority students in introductory physics

A dissertation presented by

by

Jessica Ellen Watkins

to

the School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Applied Physics

Harvard University

Cambridge, Massachusetts

May 2010

© 2010 by Jessica Ellen Watkins

All rights reserved.

Examining issues of underrepresented minority students in introductory physics courses

Eric Mazur

Jessica Watkins

Abstract

In this dissertation we examine several issues related to the retention of under-represented minority students in physics and science. In the first section, we show that in calculus-based introductory physics courses, the gender gap on the FCI is diminished through the use of interactive techniques, but in lower-level introductory courses, the gap persists, similar to reports published at other institutions. We find that under-represented racial minorities perform similar to their peers with comparable academic preparation on conceptual surveys, but their average exam grades and course grades are lower. We also examine student persistence in science majors; finding a significant relationship between pedagogy in an introductory physics course and persistence in science.

In the second section, we look at student end-of-semester evaluations and find that female students rate interactive teaching methods a full point lower than their male peers. Looking more deeply at student interview data, we find that female students report more social issues related to the discussions in class and both male and female students cite feeling pressure to obtain the correct answer to clicker questions.

Finally, we take a look at an often-cited claim for gender differences in STEM participation: cognitive differences explain achievement differences in physics. We examine specifically the role of mental rotations in physics achievement and problem-solving, viewing mental rotations as a tool that students can use on physics problems. We first look at student survey results for lower-level introductory students, finding a low, but significant correlation between performance on a mental rotations test and

performance in introductory physics courses. In contrast, we did not find a significant relationship for students in the upper-level introductory course. We also examine student problem-solving interviews to investigate the role of mental rotations on introductory problems.

Table of Contents

<i>Abstract</i>	<i>iii</i>
<i>Table of Contents</i>	<i>v</i>
<i>List of Figures</i>	<i>viii</i>
<i>List of Tables</i>	<i>ix</i>
<i>Acknowledgements</i>	<i>xi</i>
Chapter 1: Introduction	1
Under-representation of females and racial and ethnic minorities.....	1
Reforms in physics education.....	3
Pedagogy and under-represented minority students.....	4
Impact of PER reforms on retention in science majors.....	4
Impact of PER reforms on learning and achievement in introductory courses.....	5
Response to PER reforms.....	6
Further research in understanding gender and race disparities on physics.....	7
Organization of the dissertation.....	8
Chapter 2: Interactive teaching and under-represented students' achievement and retention	10
Introduction.....	10
Background.....	11
Peer Instruction.....	11
Reformed teaching and gender differences in learning and achievement.....	21
Reformed teaching and differences by race and ethnicity.....	28
Persistence and pedagogy.....	33
Research questions.....	35
Methods.....	36
Description of courses.....	36
Measures.....	39
Sample.....	42
Analytic methods.....	45
Results.....	49
Effects of gender and race on students' conceptual understanding.....	49
Effects of gender and race on students' course achievement.....	60
Retention in the sciences.....	71
Discussion.....	76
Gender differences reduced in some IE2 courses.....	76
Racial differences found on exams, grades, but not conceptual surveys.....	77
Interactive teaching linked to student persistence in STEM majors.....	80

Chapter 3: Gender, race, and student ratings of instruction	82
Introduction	82
Background	84
Are student evaluations of teaching related to student learning?	84
Student ratings of instruction are multidimensional.....	89
Student ratings of instruction and their beliefs about learning.....	92
Student perceptions of interactive teaching.....	95
Methods.....	97
Course information.....	97
Measures.....	98
Sample	112
Analytic methods.....	114
Results	117
Relationship between student ratings of interactive teaching and learning	117
Relationship between student ratings and students' expectations about physics	123
Race and gender differences in student ratings of interactive teaching	124
Discussion	132
Chapter 4: Understanding female students' responses to interactive teaching	135
Introduction	135
Background	136
Stressful collaborations	136
Gender differences in discourse	138
Inequity in participation	139
Methods.....	142
Interview participants	142
Interview data	143
Analytic methods.....	145
Results	146
Social issues contribute to differing experiences for males and females	147
Some male and female students perceive a 'testing' atmosphere during PI and respond negatively to the stress of answering ConcepTests correctly	157
Discussion	162
Conclusion.....	167
Chapter 5: Gender, mental rotations, and introductory physics	168
Introduction	168
Background	169
What cognitive sex differences have been documented?	169
What are mental rotations and how are they assessed?.....	170
What is the relationship between mental rotation and science learning?	175
Conceptual framework	179
Methods.....	184
Quantitative data.....	186
Qualitative data.....	193
Survey results	199

Gender differences in performance on MRT, CSEM, and course achievement	199
Relationship between MRT and course achievement	201
Results from student interviews	204
Describing the tools.....	204
Examining and quantifying the tools students.....	227
Case studies from student interviews	233
Discussion	249
Chapter 6: Conclusions and implications for teaching and learning.....	256
Summary of conclusions	256
Achievement and retention of under-represented minorities in PI courses.....	256
Student response to interactive teaching: gender differences.....	258
Sex differences on mental rotation tasks and introductory physics	259
Insights from an apprenticeship perspective of learning.....	260
Learning as becoming a part of a community	261
Applying these ideas to the classroom	262
Implications of this perspective for our results on under-represented minorities	264
Reflections on discipline-specific education research	268
References.....	269
Appendix 1: Mechanics Baseline Test results	290
Appendix 2: Pre-course and post-course surveys for student expectations and ratings.....	291
Appendix 3: Interview protocols on student expectations and response to PI.....	301
Appendix 4: Interview problems for mental rotations chapter.....	308

List of Figures

2.1	An example of a ConcepTest.....	12
2.2	Peer Instruction implementation procedure.....	13
2.3	Normalized gain on FCI, by pedagogy.....	16
2.4	Average CSEM scores, by pedagogy.....	18
2.5	Average FCI scores, by gender and pedagogy.....	23
2.6	Comparison of males' and females' fitted odds of scoring high on the FCI.....	54
2.7	Average racial differences on CSEM, final exam and final grade in Physics 1b..	69
2.8	Percentages of students switching out of science majors by year.....	73
2.9	Fitted probabilities of switching out of science.....	75
3.1	Histogram of student ratings of the teaching methods used in Physics 11b.....	122
3.2	Scatterplots and fitted plots for students' expectations versus ratings.....	131
5.1	Sample item from the redrawn Vandenberg-Kuse Mental Rotation Test.....	171
5.2	Sample item from the Purdue Visualization of Rotations Test.....	172
5.3	Depiction of the interactions in mediated action.....	182
5.4	Problem 3 from talk-aloud interviews.....	194
5.5	Screenshot of student using hands to build physical representation.....	197
5.6	Screenshot of student referring to picture while doing right-hand rule.....	218
5.7	Student diagram rotating the axes and showing the B-field and wire.....	218
5.8	Student using pen and hands to model the axes and wire.....	220
5.9	Student using hands to represent the rotation of the wire without looking at it..	221
5.10	Tool use among students who answered problem 3b correctly.....	229
5.11	Tool use among students who answered problem 3b incorrectly.....	231
5.12	Maverick's drawing on diagram to represent the rotation of the wire.....	235
5.13	Kyle's diagram representing the rotation of the wire in the xy plane.....	238
5.14	Hawk's graph of magnitude of the force versus rotation of the wire.....	246

List of Tables

2.1	Description of courses in this study	38
2.2	Student demographics and majors in all courses	43
2.3	Comparison of students with no missing data and students with missing data	44
2.4	Fitted logistic regression models predicting FCI high scores	51
2.5	Fitted regression models predicting Physics 11b CSEM scores	58
2.6	Fitted regression models predicting Physics 1b CSEM scores	59
2.7	Fitted regression models predicting final exam scores in 11a and 1a	61
2.8	Fitted regression models predicting Physics 1b final exam scores	64
2.9	Fitted regression models predicting final grades in Physics 11a and 1a	67
2.10	Fitted regression models predicting Physics 11b final grades	68
2.11	Fitted regression models predicting Physics 1b final grades	68
2.12	Racial breakdown of non-exam grades in Physics 1b	71
2.13	Percentages of students switching out of science, by pedagogy and gender	72
2.14	Fitted regression models predicting odds of switching out of science majors	74
3.1	Clusters of items from the precourse survey administered in Physics 11b	104
3.2	Items included in <i>methods</i> cluster	111
3.3	Student demographics and majors	112
3.4	Response rates on conceptual surveys and student evaluations	113
3.5	Comparison of students with no missing data and students with missing data	114
3.6	Univariate summary statistics on students' survey data and grades	120
3.7	Estimated correlations relating students' SETs to course performance	120
3.8	Estimated correlations relating students' ratings to course performance	122
3.9	Estimated correlations relating students ratings to their expectations	124
3.10	Gender differences on student ratings, expectations, and course performance	126
3.11	Fitted regression models predicting student ratings of the methods	128
3.12	Multiple imputation estimates in regression models predicting student ratings	128
4.1	Background information for interviewees	144
4.2	Social issues and testing themes that emerged from the interviews	146
5.1	Response rates on MRT and conceptual surveys	188
5.2	Comparison of PS students with matched data and students with missing data	190
5.3	Comparison of HP students with matched and students with missing data	190
5.4	Comparison of male and female PS students based on MRT participation	174
5.5	Survey scores and course grades for interview participants	195
5.6	Univariate summary statistics of PS MRT and course performance by gender	200
5.7	Summary statistics of HP MRT and course performance by gender	200

5.8	Estimated correlations relating MRT scores to course performance	202
5.9	Multiple imputation estimates in regression models predicting PS exams.....	203
5.10	Multiple imputation estimates in regression models predicting HP exams.....	203
5.11	Students answering problem 3b correctly, by course and MRT score	228
5.12	Low MRT and high MRT students' tool usage on problem 3b.....	228
A1.1	Fitted multilevel regression results predicting MBT scores	290

Acknowledgements

When I think about my decision to go to graduate school, I cannot believe how little I knew about what this journey would be like. I had no idea that I would meet so many different people who would shape my experiences, challenge my ideas, and become my dear friends. These people, as well as many other family, friends, and teachers prior to graduate school, are as much a part of this thesis as I am. There are so many people I must thank for helping me reach this point in my life that if I included them all, my thesis would have to be in two volumes.

First, I am so incredibly grateful to have both my mom and my dad supporting me throughout my life. They are very different people, but I like to think that the two different perspectives I got from each of them grounded me more deeply. I want to especially thank my loving mother for emphasizing the value of education and always encouraging me to stand up for the things I care about. As for my dad, his unconditional support and generosity throughout my life allowed me to reach farther and try new things.

I am very lucky to have a wonderful family who guided me growing up and encouraged me from afar ever since. My stepfather Harold not only proofread almost all of my essays in secondary school, but he taught me how to work hard in his example. While my stepmother Annie is a newer addition to my family, I admire her patience and independent spirit. My brother Blake has always been a good friend and I appreciated his

support especially during the writing of this dissertation. Finally, my sister Clara is the most imaginative, kindest little girl I have ever known—her performances always lifted my spirits during my visits home.

I can honestly say that this degree would not have been possible without the help and support of Eric Mazur. He helped to carve a place for me and my research both in his group and in the department. He was supportive of my professional life by allowing me to pursue my research interests, encouraging and supporting me to visit other research groups, and helping me overcome the hurdles that arose in my work. Beyond the research however, Eric is an advisor who truly cares about his students' lives outside of lab and I greatly appreciate that.

Early in my graduate school career I met Eugenia Etkina, whose guidance and wisdom were central in my development as a scientist. She was an integral advisor in the conception and planning of most of the projects in this thesis and offered unwavering support throughout the research. She has also become a dear friend and personal mentor—without her this process would not have been nearly as rewarding.

I also must thank my committee members, John Willett, Marie Dahleh, and Michael Aziz, for supporting my research and providing advice whenever asked. In particular, this work is greatly shaped by my meetings with John and his intellectual generosity and thoughtful criticism. His guidance was invaluable not only for my research, but also in thinking about my career plans. I also truly appreciate the time he spent reading and editing this dissertation.

As much as I learned from the professors I interacted with at Harvard, I learned just as much from my friends in the Mazur group. The people here, both past and present,

showed me what a scientific community is. I appreciate the patience that Rafael, Jim, Maria, Iva, Brian, Sam, Geoff, Paul, Tobias, and Tommaso all displayed while teaching me about science, graduate school, and life. I valued learning alongside Prakriti, Loren, Tina, Mark, Eric, Kevin, Jason, Chris, Mustafa, Renee, Sally, Yuting, Michael, and Kasey. In particular, I shared most of these past six years with Eric and Mark, whose insight, humor, and kindness were much appreciated.

While the education subgroup was often quite small, I benefited greatly from the many visitors who joined us; Mercedes, Jessica Winkler, and Katja all helped with various aspects of this thesis. I also am extremely appreciative for the expertise and guidance of Veronica, Jessica Rosenberg, Nathaniel, and Julie; they shared their perspectives on research and encouraged my intellectual development. Vijay, Martin, and Doug also were all instrumental in keeping the technology of the subgroup running. As a Mazur group alumna, Catherine Crouch was influential on much of the gender work presented here and I am thankful for her comments on one of the chapters. Finally, I am especially grateful to Jason, who made significant intellectual contributions to the mental rotations chapter and continues to be thoughtful colleague and friend.

Beyond the Mazur group, there were many others at Harvard who not only helped with this thesis, but enriched my time in graduate school. Simone Pinheiro was a phenomenal housemate and friend, who was always willing to listen to me talk about my work even if she was buried in her own. I enjoyed collaborating with fellow students, staff, and researchers at the Bok Center. In particular, Erin Driver-Linn, Cassandra Volpe-Horii and John Girash were excellent mentors and were quick to provide excellent advice and support. Karen Pearce was not only helpful in procuring data, but in

discussing the methods and results of the quantitative analysis. I also thank Logan McCarty and Masahiro Morii for allowing me to collect data in their courses and for their many useful suggestions. Finally, I am indebted to the students in this thesis for their willingness to participate and share their ideas about physics.

Throughout the last several years I have been privileged to expand my scientific community beyond Harvard and my committee. My interactions with other physics education researchers inspired and enhanced my research and thoughts about learning physics. In particular, I am indebted to the PER@Colorado group for hosting me and offering constructive feedback on my work. I am especially appreciative for the insights of Chandra and Lauren. Furthermore, members of the UMD PERG were willing to listen and provide thoughtful feedback from an early stage in my research career—particularly Joe Redish and Andy Elby. More recently, I am grateful for the support of new UMD friends and colleagues as I finished this dissertation. Renee Michelle was an excellent writing partner and friend. I am also indebted to Joe for offering helpful suggestions on several chapters of this thesis.

I am also very thankful to all of my teachers over the past 23 years of my education. From an early age I found the gift of argumentation and I especially appreciate the teachers who not only put up with that, but supported and encouraged independent thinking. In particular, I want to thank Mr. Shea, Mr. Jaffurs, Mr. Miller, Hutch, Dr. Colvin, and Dr. Bieclazyc—your personal attention had an impact on me, so thank you.

Finally, I cannot go without thanking my best friend Jared. His love and encouragement were always there for me, even if he was physically hundreds or thousands of miles away. He has supported me through every phase of my graduate

school life, celebrating with me in my successes and providing me strength during my trials. My life is happier, more colorful, and richer because of his presence.

*Jessica Watkins
Cambridge, MA
May, 2010*

Acknowledgements of Financial Support

This thesis is based on work funded by the National Science Foundation under contracts DUE-0206947 and DUE-0716902, the Davis Foundation, and the Engineering Information Foundation.

Chapter 1

Introduction

1.1 Under-representation of females and racial and ethnic minorities in physics

In 1962, an article concerning the under-representation of women in physics appeared in the *American Journal of Physics* (Clancy, 1962). Emphasizing the differing social pressures that males and females faced, Clancy discussed many possible explanations for the disparity in participation at the upper levels of the discipline. In the ensuing decades, dramatic strides were made in the participation and success of females in physics, science and engineering, and higher education in general (NSF, 2008). For example, since 1982, the majority of students attending U.S. colleges and universities have been female. Furthermore, female students earned about half of the science and engineering bachelor's degrees in the last ten years (NSF, 2010). However, although the percentage of females earning undergraduate and graduate degrees in physics has increased, the discipline still lags behind the other natural sciences. Data from 2007 show that while female students earn 60% of biological-science bachelor's degrees and 50% of chemistry degrees, they earn only 21% of the physics bachelor's degrees (NSF, 2010). Similarly, in 2006, half of the science and engineering graduate students were female, but only 20% of physics

graduate students were female, which is less than all other science, technology, and engineering (STEM) disciplines (NSF, 2010).

The problem of female under-representation in physics is often considered in terms of a “leaky pipeline,” in which disproportionate percentages of women leave the field at every step of the academic ladder. However, recent analyses show that this metaphor may be inappropriate. “Indeed, the data show that there is only one significant leak in the physics pipeline, and that is between taking high-school physics and earning a bachelor’s degree in physics.” (p. 13, Ivie & Ray, 2005) Therefore, understanding what happens to females between high school and college graduation is imperative for addressing the gender disparity in the participation and success in physics.

Unlike females, the under-representation of some racial and ethnic minority groups is a much broader problem, not limited to physics nor even STEM fields. Although the percentages of 25–29 year-olds with a bachelor’s degree increased for all three racial groups between 1971 and 2007, so did the gap between Whites and their Black and Hispanic peers (NCES, 2008). In 2007, under-represented racial- and ethnic-minority students, which the National Science Foundation classifies as Black, Hispanic, and American Indian/Alaskan Native students, earned 18% of all bachelor’s degrees (NSF, 2010). In science and engineering, 17% of bachelor’s degrees were awarded to under-represented minorities. However, the problem is even worse in physics: less than 9% of the physics bachelor’s degrees were earned by under-represented racial- and ethnic-minority students in 2007. Even lower percentages (6–7%) of graduate students in physical sciences were under-represented minorities in 2006.

1.2 Reforms in physics education

One avenue to address under-representation of these minority groups is to focus on how physics courses are taught at the undergraduate level. In the past few decades, great advances have been made in understanding how to improve physics instruction, particularly at the introductory level. Physics education research (PER) has provided insight into student thinking in physics in these courses. The PER community has examined a broad range of issues, including students' understanding of physics concepts, how students can apply intuitive ideas to build their physics knowledge, what students think it means to learn physics, and students' attitudes and expectations while learning physics. The results of this research have informed classroom teaching in physics courses across the country (for a review, see Thacker, 2003). In reformed classrooms, the physics textbooks that are often used were written to incorporate results on student difficulties with specific physics concepts (R. D. Knight, 2007; Mazur, in print); homework and exams are redesigned to address more sophisticated learning goals for the students (e.g., Mazur, 1997); and innovative pedagogies are incorporated that require the students to interact with their peers and actively think about physics concepts during class (e.g., Beichner & Saul, 2003; Etkina & Van Heuvelen, 2007; Mazur, 1997; McDermott, Schaffer, & UWPERG, 2002; Redish, 2003). The novel pedagogies born out of PER have greatly improved the teaching and learning of physics in both introductory- and advanced-level courses. In particular, the interactive teaching methods have been shown to improve students' conceptual understanding in introductory-physics courses in diverse settings across the U.S. (Hake, 1998). However, the impact of these new teaching methods on the under-representation of females and racial- and ethnic-minority students

is less well-known; more research is needed on how under-represented minority students learn in these courses.

1.3 Pedagogy and under-represented minority students

While PER reforms have benefitted students in diverse settings and for myriad learning goals, much less is known about their impact on under-represented minorities. In this section, we discuss what research needs to be initiated or expanded to better understand how instructors and institutions can address the under-representation of females and racial- and ethnic-minority students in physics.

1.3.1 Impact of PER reforms on retention in science majors

While much research has been conducted on the impact of PER reforms on *all* students' performances on conceptual surveys and exams, much less has been discussed on whether the effects of exposure to these new teaching methods predicts student persistence in physics or *all* science, engineering, and technology (STEM) majors. In Sheila Tobias's (1990) book, *They're Not Dumb, They're Different*, she asked post-graduate students with considerable academic preparation to audit a chemistry or calculus-based physics course to provide a unique perspective of students' experiences in these introductory courses. She found that the pedagogies used in these traditional lecture courses turn these students off, particularly because of the isolating, competitive environments, didactic teaching methods, and lack of emphasis on understanding concepts. Tobias rightly advocates for a change in introductory courses to help retain students who do not respond well to the traditional teaching methods. As PER reforms

have addressed many of the issues that students in Tobias's book discuss, it is surprising that there is such sparse data in the PER literature on student persistence. To address whether teaching methods can have impact on the under-representation of females and racial and ethnic minorities, there needs to be more focus on the long-term impacts for all students as well as under-represented students.

1.3.2 Impact of PER reforms on learning and achievement in introductory courses

After examining the impact of pedagogy on student retention, it is reasonable to next look at gender and racial differences in student performance in introductory courses, as these results will further inform investigations of the under-representation of minority students in physics. These courses, which are often taken at the start of students' postsecondary education, provide the first feedback to students about how well they can succeed in the field. Students who receive lower exam scores and grades in these courses may decide that they are not well enough prepared or able to major in physics. Furthermore, final grades have significant external impacts on their careers—affecting whether students can continue in a major or get accepted to postgraduate school.

In addition to student grades, looking at low-stakes assessments is useful to understanding the under-representation of minority students. Much of the PER reform movement was initiated, and is driven by, assessment of introductory students' conceptual understanding through the use of standardized surveys (Beichner, 2009), such as the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992) or Conceptual Survey of Electricity and Magnetism (Maloney, O'Kuma, Hieggelke, & Van Heuvelen, 2001). While students may not receive their posttest scores, differences in performance

on these assessments may have a lasting impact on under-represented students' persistence in the major. Students who perform poorly may not understand the basic concepts in introductory physics, which will pose great challenges as they take more advanced courses which build on these concepts. Also, as students take the assessment, they may recognize that they do not fully understand the physics in their courses, providing less concrete, but still valuable feedback about whether they should continue in the physics major.

The bulk of the work done in physics education on the under-representation of female students focuses on the differences in performance on conceptual surveys and exams, often examining the impact of reformed pedagogy on the gender gaps in introductory physics. As discussed in Chapter 2, the results to date are inconclusive, warranting further research on whether interactive, constructivist, student-centered teaching methods can reduce or eliminate gender differences in performance on conceptual surveys, exam scores, and final grades. Much less work has been done at this level on under-represented ethnic and racial minorities—while analyses of student grades suggests that there are racial differences (Hazari, Tai, & Sadler, 2007), little is known about differences in performance on conceptual surveys or how reformed instruction may mediate if any differences are observed.

1.3.3 Response to PER reforms

While most of the PER research on the under-representation of females in physics examines gender differences in learning and achievement, it is likely that students' feelings about and response to the instruction are equally important in whether they

pursue STEM (or, more specifically, physics) as a major and career. As discussed above, Sheila Tobias found that some well-prepared, motivated, and successful post-graduate students responded negatively to the traditional teaching methods used in introductory courses, suggesting that response to instruction may impact retention in the sciences. With the introduction of new teaching methods, further investigations are necessary to understand how students, particularly under-represented female and racial- and ethnic-minority students, experience and respond to novel ways of teaching this material.

1.3.4 Further research in understanding gender and race disparities in physics

In addition to examining the impact of novel teaching methods on under-represented minority students, it is worthwhile to conduct additional empirical research to better understand the nature of gender (and race) differences and to inform further pedagogical reforms. In particular, much has been said and written about the nature of the differences between males and females in participation and success in the physical sciences. Many of the same issues brought up in the 1962 *American Journal of Physics* article are still being discussed today—differing social pressures, aptitudes, preferences, and status. Further research can inform (and has informed) debates on the gender and racial disparities as well as direct policy on these issues. For example, research in psychology revealed that stereotype threat, the risk of confirming a negative stereotype about a group with which a student identifies, can depress students' test performances (Steele & Aronson, 1995). Results confirming the effects of stereotype threat have been obtained for both women and racial and ethnic minorities in science and mathematics (Spencer, Steele, & Quinn, 1999; Steele & Aronson, 1995). These findings have sparked

further research on the bias of standard measures of academic performance as well as on how to design interventions to reduce the stereotype-threat effects in school settings (Walton & Spencer, 2009). Currently, PER researchers at University of Colorado, Boulder are investigating the effects of stereotype threat to determine the impact in introductory physics courses. Further research is needed on this issue and the many other cognitive, psychological, and social issues proposed to explain gender and race differences in physics.

1.4 Organization of the dissertation

In this dissertation, we address each of the above issues, which all relate to the disproportionate participation and success of female and under-represented racial- and ethnic-minority students. We focus on students in reformed and traditionally-taught (lecture-based) courses at Harvard University for our analyses. Given that Harvard University is a highly selective Research I institution, our results may not extend to the general population of introductory physics students. However, results concerning this student population may be an especially useful first step in examining under-representation of females and particular racial and ethnic minorities in graduate programs; students at selective undergraduate institutions are the most likely to pursue advanced graduate degrees (Mullen, Goyette, & Soares, 2003).

In Chapter 2, we extend previous analyses on the relationship between interactive teaching and gender differences in introductory physics courses. We also examine the performance of under-represented racial- and ethnic-minority students in these reformed courses. Finally, we address the relationship between interactive teaching and student

persistence in STEM majors, highlighting the long-term impact of pedagogy in a single course.

In Chapter 3, we examine survey data on student response to the interactive pedagogies used in introductory physics courses, finding that females respond more negatively to the teaching methods.

In Chapter 4, we further investigate the gender differences found in Chapter 3, focusing on student interviews and previous research to hypothesize why these female students experience and respond more negatively to the interactive teaching methods used in this course.

In Chapter 5, we explore the research on biological sex differences and examine one often-cited link between performance on tests of mental rotations and the under-representation of females in physics.

In Chapter 6, we summarize the findings and place them in the context of current thinking in educational theory, exploring the role of community and apprenticeship in under-represented students' learning.

Chapter 2

Interactive teaching and under-represented students' achievement and retention

2.1 Introduction

Publications from the PER community show that interactive, student-centered pedagogies in introductory physics courses can improve students' conceptual understanding as measured through written assessments and student interviews (e.g., Crouch & Mazur, 2001; Hake, 1998; Shaffer & McDermott, 1992; Sokoloff & Thornton, 1997). Although this research demonstrates that the average learning gains for all students can be improved, more research is needed about the impact of these teaching interventions on under-represented minority students. Recently, researchers have studied whether interactive pedagogies reduce gender differences (e.g., Kost, Pollock, & Finkelstein, 2009; Lorenzo, Crouch, & Mazur, 2006; Pollock, Finkelstein, & Kost, 2007); however, very little is known about how interactive teaching impacts under-represented racial and ethnic minorities. Furthermore, most research on increasing the participation of under-represented racial minorities in STEM focuses on more comprehensive interventions at the institutional or departmental level, not on pedagogical interventions at the course-level (Olson, Fagen, & National Research Council Board on Life Sciences, 2007).

In this chapter, we take a closer look at the impact of Peer Instruction and other interactive teaching techniques on under-represented minority students at Harvard University. We re-analyze data comparing the conceptual performance of male and female students, using multiple regression analysis to control for differences in students' background preparation. Furthermore, we expand our analysis to investigate differences in course performance, again using multiple regression analysis. We also present the first results examining the effects of interactive teaching on under-represented racial- and ethnic-minority students, looking at differences in performance on conceptual inventories, course exams, and grades. Finally, we investigate the long-term impact of interactive teaching on student persistence in STEM majors.

In the next section, we describe in detail Peer Instruction (PI), the primary teaching intervention under investigation in this chapter. We also review previous research on reducing the gender and race gap in achievement in science courses, focusing on the impact of pedagogy. Finally, we detail relevant research in student persistence in STEM majors, which highlights the impact of poor teaching on students leaving science at the postsecondary level.

2.2 Background

2.2.1 Peer Instruction

What is Peer Instruction?

PI is an interactive teaching technique that promotes classroom interaction to engage students and address difficult aspects of the material (Crouch & Mazur, 2001;

Crouch, Watkins, Fagan, & Mazur, 2007; Mazur, 1997). PI structures time during class around short, conceptual multiple-choice questions, known as ConcepTests. In Figure 2.1, we show an example ConcepTest from introductory physics. These questions are targeted to address student difficulties and promote student thinking about challenging concepts.

Figure 2.1. An example of a ConcepTest.

A blood platelet drifts along with the flow of blood through an artery that is partially blocked by deposits.



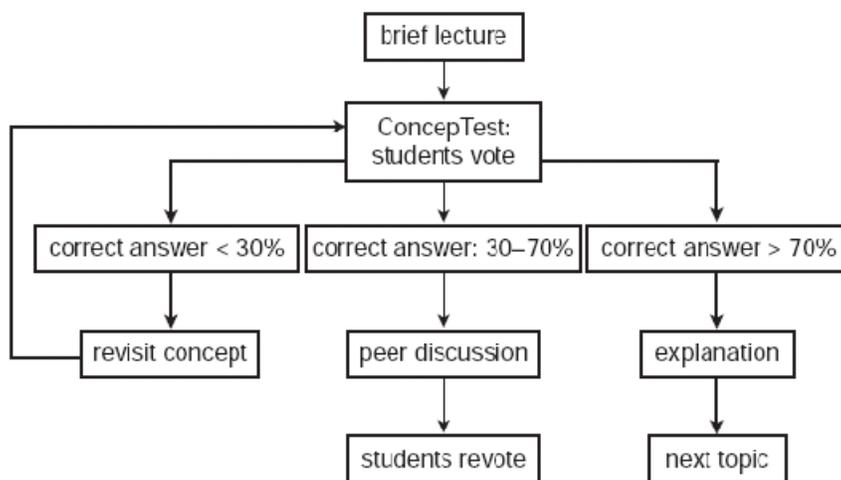
As the platelet moves from the narrow region to the wider region, its speed

1. increases.
2. remains the same.
3. decreases.

The ConcepTest procedure is depicted in Figure 2.2. After a brief presentation by the instructor, the focus shifts from instructor to student, as the instructor encourages students to think about the material by posing a ConcepTest. After 1–2 minutes of thinking, students commit to an individual answer. If too few students respond with the correct answer, the instructor may revisit the concept using lecture or try a different ConcepTest. If a large majority of students responded correctly, the instructor typically gives a brief explanation and moves on the next topic or ConcepTest. If an appropriate percentage of students answer the ConcepTest correctly, the instructor asks students to turn to their neighbors and discuss their answers. Students talk in pairs or small groups and are encouraged to find someone with a different answer. The teaching staff circulates

throughout the room to encourage productive discussions and guide student thinking. After several minutes, students answer the same ConcepTest again. The instructor then explains the correct answer and, depending on the student answers, may pose another related ConcepTest or move on to a different topic.

Figure 2.2. A Peer Instruction implementation procedure



In science courses, PI has been shown to be a useful way to engage students in classroom demonstrations, much like Interactive Lecture Demonstrations (Sokoloff & Thornton, 1997). Before showing students what happens when you mix two chemicals or flip a switch on a given circuit, instructors can ask students to predict the outcome. Research shows that asking students to predict the outcome of an experiment results ultimately in greater conceptual understanding (Crouch, Fagen, Callan, & Mazur, 2004), and instructors also report increased student engagement (Mazur, 1997). In social-science or humanities courses, PI can be used to involve the students as participants in experiments with human responses (S. W. Draper, Cargill, & Cutts, 2002).

A variety of question-types can be used with PI, including questions about general theories and definitions, questions that ask students to apply concepts in different contexts, and questions that illustrate how different ideas are related. PI is not only useful for questions with “correct” answers, but also for promoting discussion among students with questions that lack a clear-cut answer. For example, a ConcepTest may ask students to consider the relative importance of different assumptions in a scientific hypothesis or the relative value of different interpretations of a literary passage. The structure of PI provides opportunities for students to hone their skills in critical listening and developing solid arguments. Regardless of subject matter, PI promotes a learning environment in which students to create knowledge through discussion and become active participants in the discipline they are studying.

The quality of student discussion and learning in a PI classroom depends on the quality of the ConcepTests. Several databases of class-tested questions exist in physics (Mazur, 1997), chemistry (Ellis, et al., 2000; Landis, et al., 2001), astronomy (Green, 2002), mathematics (Hughes-Hallet, et al., 2003; Terrell, 2005), geoscience ("Starting Point, Teaching Entry Level Geoscience: ConcepTest Examples," 2008), philosophy (Bigelow, Butchart, & Handfield, 2007), and psychology ("Canadian In-Class Question Database: Psychology," 2005). For a ConcepTest to be most effective, the question must require higher-level thinking about a concept so that students aren't simply recalling something they read or using “plug-and-chug” with equations. Questions must also be at an appropriate difficulty level so students are challenged but can reason to the answer with their existing knowledge.

Previous research indicates that students get the most benefit from peer discussion when about 30–70% of the class answers the ConcepTest correctly before discussion (Mazur, 1997). Too few correct answers may indicate that students do not have enough understanding nor knowledge to engage productive discussions. Therefore, students must come to class with some knowledge and ideas about the material. Often instructors administer reading quizzes at the start of class to promote pre-class reading; however, this assignment often relies solely on student memorization of facts, definitions, or equations. Another method, Just-in-Time-Teaching (JiTT; Novak, Gavrin, Christian, & Patterson, 1999; Novak, Patterson, Gavrin, Christian, & Forinash, 1999; Simkins & Maier, 2009) encourages students to read the material by asking students to answer two short-answer questions about the reading before class. These questions often ask for more than memorization of key words and definitions and push students to start thinking more deeply about the concepts. These assignments are typically completed online, so instructors can review students' answers while planning their lectures.

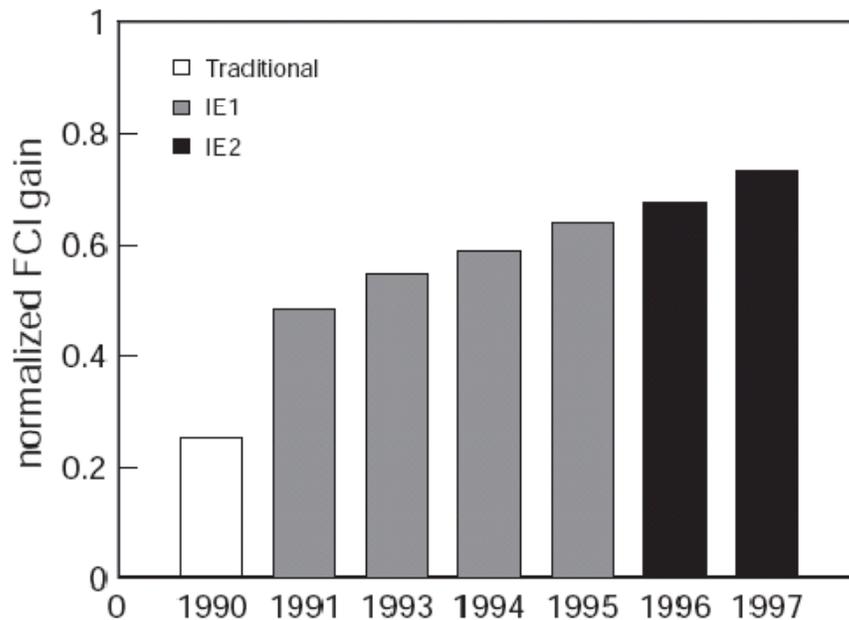
Research shows that PI works

Research in physics education shows that courses incorporating “activities that yield immediate feedback through discussion with peers and/or instructors” (p. 65) result in greater student conceptual understanding than traditional courses (Hake, 1998). Data from introductory physics courses at Harvard University confirm this finding for PI, as seen in Figures 2.3 and 2.4. In Figure 2.3 we compare results from a traditional course and several PI courses using a standardized conceptual assessment of Newtonian mechanics, the Force Concept Inventory (FCI; Hestenes, et al., 1992). As a measure of student learning, the average normalized gain (Hake, 1998) was estimated for each

course, which represents the gain from pretest to posttest, divided by the maximum gain possible (100% minus pretest score): $g = \frac{post - pre}{100 - pre}$. PI courses (IE1) obtained greater

learning gains than traditional courses. In 1996 and 1997, reading assignments and small-group activities, including Tutorials in Introductory Physics (McDermott, et al., 2002), were used with PI (IE2), which resulted in even higher normalized gains (Crouch & Mazur, 2001). This combination of several interactive, feedback-intensive methods received the highest gains in student performance on a conceptual test.

Figure 2.3. Normalized gain on the Force Concept Inventory in calculus-based introductory physics courses at Harvard University, by pedagogy: traditional course (1990), IE1 courses which used PI (1991, 1993-1995), and IE2 courses which used reading assignments, PI, and other interactive techniques (1996-1997). (reprinted from Crouch & Mazur, 2001)



The increased overall learning gains with the use of PI occurred not only at Harvard University. The results were replicated at a community college (Lasry, Mazur, &

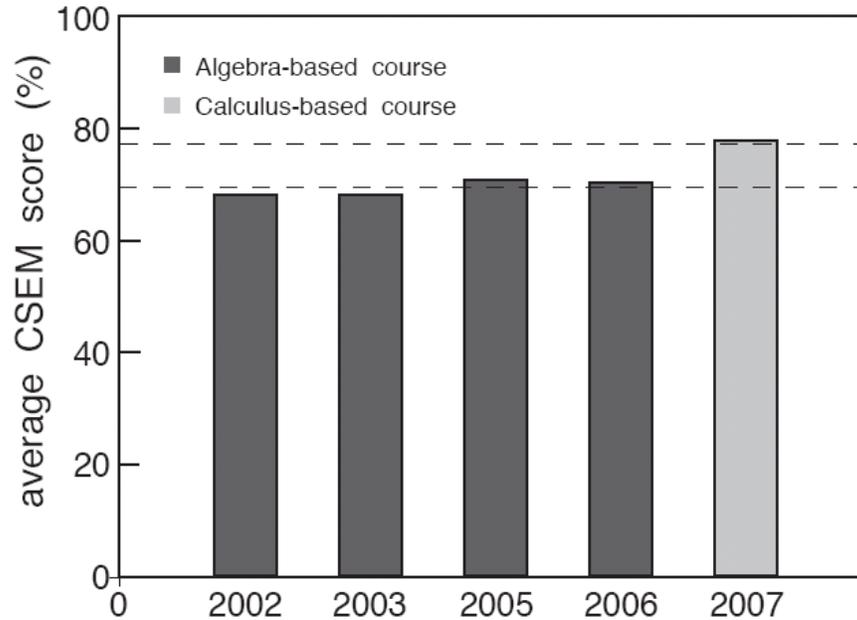
Watkins, 2008), suggesting that PI was effective with heterogeneous student populations. Additionally, the positive results of PI are not limited to physics courses. Other studies have shown that PI is useful in improving learning in biology (J. K. Knight & Wood, 2005), engineering (Nicol & Boyle, 2003), psychology (Morling, McAuliffe, Cohen, & DiLorenzo, 2008), medicine (Rao & DiCarlo, 2000), philosophy (Bigelow, Butchart, & Handfield, 2006), and mathematics (R. L. Miller, Santana-Vega, & Terrell, 2006).

In Figure 2.4, we show the results from introductory electricity and magnetism courses with the Conceptual Survey of Electricity and Magnetism (CSEM; Maloney, et al., 2001). Each of these courses incorporated PI, JiTT, Tutorials and cooperative group problem-solving exercises. Students in algebra-based non-major physics courses achieved average scores similar to those obtained by senior undergraduate physics students, and students in the calculus-based course achieved an average score comparable to that of two-year college physics professors (Maloney, et al., 2001; Watkins & Mazur, 2009).

Why PI works

A great deal of research on cognition and learning indicates that students learn by using their existing knowledge, beliefs, and skills to create new knowledge (see Bransford, Brown, & Cocking, 2000 for a review). Therefore, pedagogies in which instructors are made aware of students' incoming knowledge can enhance learning. PI provides opportunities for instructors and students to recognize background knowledge during the pre-class reading, their initial vote, and discussion. The best in-class ConcepTests often take advantage of the experiences and thinking students bring to the classroom about the material so that students can recognize their ideas and build on them.

Figure 2.4. Average scores on the Conceptual Survey of Electricity and Magnetism for algebra-based and calculus-based introductory physics courses. Dotted lines represent the average scores obtained by advanced undergraduate physics majors (bottom line) and 2-year college professors (top line). (reprinted from Watkins & Mazur, 2009)



With the constant feedback from the reading assignments and ConcepTests, the instructor can monitor student progress and help guide students to use their previously-held ideas to understand new concepts and theories. For example, in physics, students may not fully understand Newton's First Law because of their own conflicting experiences outside the classroom involving friction in which they slid objects on flat surfaces. However, PI helps students first express their initial ideas and then, through targeted questioning, guides them to develop more comprehensive ideas about motion that include friction. Although there are many books and papers that catalogue and describe commonly-held ideas in introductory science (e.g., Driver, Squires, Rushworth, & Wood-Robinson,

1994), JiTT is very useful in informing the instructor of these ideas before class, particularly for subjects with less research in student background knowledge. Additionally, the flexibility of a PI lecture makes it easy for instructors to spend more time on concepts that are difficult for students by giving more focused, short presentations or asking more ConcepTests. In an interactive classroom, instructors are paying attention to student thinking throughout the learning process.

With the constant feedback from the reading assignments and ConcepTests, the instructor can monitor student progress and help guide students to use their previously-held ideas to understand new concepts and theories. For example, in physics, students may not fully understand Newton's First Law because of their own conflicting experiences outside the classroom involving friction in which they slid objects on flat surfaces. However, PI helps students first express their initial ideas and then, through targeted questioning, guides them to develop more comprehensive ideas about motion that include friction. Although there are many books and papers that catalogue and describe commonly-held ideas in introductory science (e.g., Driver, et al., 1994), JiTT is very useful in informing the instructor of these ideas before class, particularly for subjects with less research in student background knowledge. Additionally, the flexibility of a PI lecture makes it easy for instructors to spend more time on concepts that are difficult for students by giving more focused, short presentations or asking more ConcepTests. In an interactive classroom, instructors are paying attention to student thinking throughout the learning process. PI provides a structured environment for students to voice their ideas and resolve misunderstandings by talking with their peers. By working together to learn new concepts and skills in a discipline, students create a more cooperative learning

environment that emphasizes learning as a community in the classroom (Hoekstra, 2008)). Research suggests that this type of cooperative learning environment can help promote deeper learning, as well as greater interest and motivation (Cross, 1998). Furthermore, the strategies students use during collaboration (explaining, reasoning, and justifying arguments) can also help students develop more advanced critical-thinking skills that can be used beyond the classroom (Gokhale, 1995).

Research also shows that experts are able to monitor and regulate their own understanding (e.g., Bransford, et al., 2000; Schoenfeld, 1992). These metacognitive abilities enable experts to employ different strategies to improve their learning. PI can help students develop better metacognitive skills, as they check their own understanding during pre-class reading and in-class questions. The method can help students recognize when they do not understand a concept, when they are unable to answer a question on the reading, or when they cannot give complete explanations to their peers during in-class discussion. With this formative, internal feedback, students can learn how to better assess their own understanding during the learning process. PI encourages students to take responsibility for their own learning and emphasize understanding over simple task completion.

PI can be adopted readily for a variety of disciplines and classroom environments, and can be modified for different instructional goals. ConcepTests can be tailored to individual classes and for diverse learning objectives. The modular nature of this method means that instructors can use PI when, and how, they see fit. Reading assignments need only be given before class when necessary, and during class instructors can use as few as one question per class or as many as time allows. As such, PI can be easily adapted to an

instructor's personal style of teaching, and combined with other teaching methods such as Tutorials (McDermott, et al., 2002), small group problem-solving (Heller, Keith, & Anderson, 1992), or lecture. It is this flexibility that makes this method so effective in so many classrooms.

2.2.2 Reformed teaching and gender differences in learning and achievement

Many of the features of PI and other PER teaching reforms are also highlighted as aspects of “gender-inclusive” pedagogy. In Brotman and Moore’s (2008) review of the literature on gender and science education, they suggest that gender-inclusive curriculum and pedagogy “includes features such as active, collaborative learning, which highlights the social relevance of science and pays particular attention to incorporating the life experiences of girls.” (p. 984) Specifically, research in both K-12 and higher education suggests that gender-inclusive teaching environments should (1) actively support cooperative learning (e.g., Baker & Leary, 1995; Labudde, Herzog, Neuenschwander, Violi, & Gerber, 2000; Laws, Rosborough, & Poodry, 1999), (2) build on *all* students’ prior knowledge, experiences, and interests (e.g., AAUW, 1992; Howes, 1998; Labudde, et al., 2000; Roychoudhury, Tippins, & Nichols, 1995), (3) decrease competitiveness (e.g., Johnson, Johnson, & Stanne, 1986; Zohar, 2003), (4) emphasize deep understanding (e.g., Hazari, et al., 2007; Tai & Sadler, 2001; Zohar, 2003), and (5) actively engage students with the material (e.g., Labudde, et al., 2000; Laws, et al., 1999; Roychoudhury, et al., 1995; Schneider, 2001). Many of these pedagogical features have been shown to benefit both male and female students, thus many results at the secondary level do not show a reduction in gender differences in science achievement through the

use of these techniques (Brotman & Moore, 2008; Labudde, et al., 2000). However, high-school courses that required a “full understanding of topics,” as opposed to “lots of memorization” benefitted females’ performance preferentially in later college physics courses (Hazari, et al., 2007).

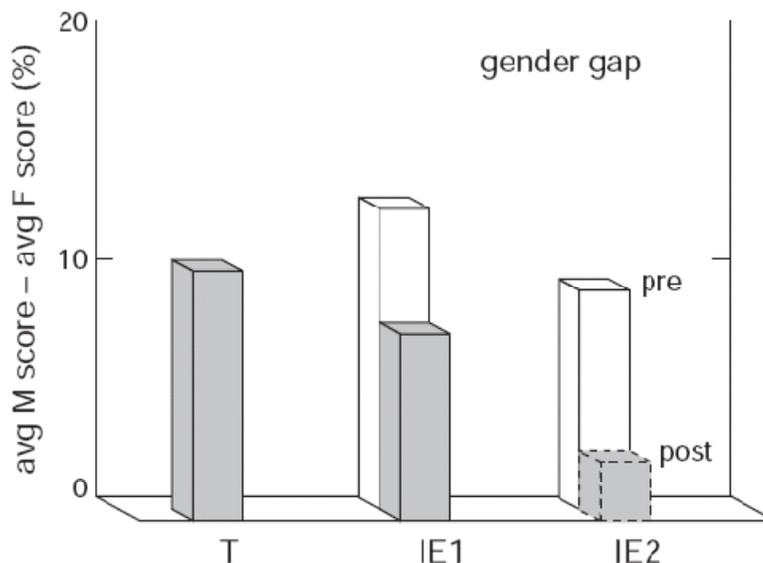
In this section, we review the research studying the impact of interactive teaching methods in introductory physics courses at the postsecondary level. We first focus on results obtained using conceptual inventories; we then examine the research on gender differences in course achievement. We find that the results published so far are inconclusive as to what impact interactive pedagogies have on the gender gap. While different institutions show different results, issues with the statistical analysis may be clouding the conclusions that we can draw from these results. We discuss these issues briefly and pose possible solutions for better examining gender differences in introductory physics courses.

Interactive teaching and gender differences on conceptual surveys

Previous work by our group suggests that PI can help diminish gender gaps in performance on conceptual tests at the postsecondary level. As seen in Figure 2.5, females entered introductory physics courses at Harvard University with lower average scores on the Force Concept Inventory than males and this gap persisted to the end of a course taught using traditional lecture methods. With just changing the lecture portion of the class using PI (IE1), the difference between males and females decreased, although the gap in posttest scores remained statistically significant. With the use of PI in combination with other interactive techniques (IE2), such as reading assignments and Tutorials in Introductory Physics, the gap in posttest scores was reduced further, until

males' and females' posttest scores were no longer different in these introductory courses at Harvard University ($p = 0.43$, Lorenzo, et al., 2006).

Figure 2.5. Differences in average scores, by gender, on the Force Concept Inventory before and after discussion with three different pedagogies: traditional, IE1, and IE2. (reprinted from Lorenzo, et al., 2006)



Further research at other institutions suggests that this reduction in the gender gap is not replicated in other student populations or with different implementations of PER methods. Researchers at the University of Colorado did not find a reduction in the gap between male and female average scores in courses taught using methods similar to both the IE1 and IE2 courses at Harvard University (Pollock, et al., 2007). These results were obtained in first-semester introductory physics courses, covering topics in mechanics, and in second-semester introductory physics courses, covering topics in electricity and magnetism. Researchers at the University of Minnesota examined the impact of using Cooperative Group Problem Solving (Heller & Hollabaugh, 1992; Heller, et al., 1992) during the small discussion sections in introductory physics courses. They found a small

reduction in the gender gap from pretest to posttest (a difference of about 2%); however the differences between males' and females' posttest scores were still statistically significant (Docktor & Heller, 2008).

Methodological issues with previous research using conceptual surveys

Much of the research described compares male and female students' posttest scores using simple statistical methods such as the application of two-tailed *t*-tests, however these methods may not be appropriate in some samples due to the failure of the normal-theory assumptions in these data. Many students, especially those in the Harvard University IE2 courses, obtained the maximum score possible on the FCI posttest, thereby creating a ceiling effect in that sample distribution of scores. Furthermore, this effect suggests that the measuring instrument is not sensitive to differences among students in the top 'bin' of scores. (Similarly, if students all obtain low pretest scores, i.e., around the percentage one would obtain by guessing, our instrument would not be sensitive to differences in background preparation.)

Previous studies also compare students' normalized and absolute gains. These two different measures of students' average improvement can lead to differing results because they represent different features of the distribution of the data (Willoughby & Metz, 2009). While use of normalized gain has been helpful for instructors to obtain a rough estimate of their students' average improvements on conceptual tests while accounting crudely for pretest scores (Hake, 1998), this approach is not appropriate for use on individual students (Marx & Cummings, 2007). Additionally, research suggests that this measure is correlated with pretest scores (Coletta & Phillips, 2005) and is biased to high pretest scores (Brogt, et al., 2007). Absolute gain, estimated by subtracting an

individual's pretest score from their posttest score, is an intuitive measure of individual improvement that can be more reliable than either the pretest or the posttest score if students are growing heterogeneously (e.g., Rogosa & Willett, 1983). However, absolute gain will also lead to misestimation of student ability and learning when data with ceiling or floor effects are used. Therefore, different analytic methods are needed to understand the effects of PI and other interactive techniques on the gender gap.

Multiple regression analysis can be used to control for differences in male and female students' background preparation, including their pretest scores. Multiple regression analysis allows us to study the simultaneous effects of multiple independent variables on an outcome variable.¹ For variables whose measurement suffers from ceiling or floor effects, we can collapse scores into cruder categorical summaries (e.g., low and high). Then, for these categorized posttest scores, logistic regression analysis can be used to predict the odds that a student will score 'high' rather than low. Previous work by Kost et al. (2009) uses these regression techniques and shows that differences in students' math preparation, beliefs about physics learning, and pretest scores predict substantial variability in students' performance on conceptual learning assessments. When controlling for these independent variables, the gender gap in performance on a conceptual posttest is no longer statistically significantly different from zero in fully interactive courses (Kost, et al., 2009).

¹ Note that linear regression is still not appropriate to use if the data exhibits floor or ceiling effects, as the distribution of the residuals will be non-normal, a violation of the assumptions needed for linear regression. Instead, we propose using logistic regression, collapsing posttest scores into categories (e.g. low and high). This analysis answers a slightly different question, however, as discussed in the methods section later in this chapter.

Gender differences in course achievement

It is also important to examine course achievement, in addition to performance on conceptual tests, as low grades can predict whether a student persists in a science major or not (Strenta, Elliott, Adair, Matier, & Scott, 1994). In high-school physics courses, females report higher grades than males (Hazari, Sadler, & Tai, 2008). At the postsecondary level, gender differences in student grades differ based on the type of introductory physics course: college physics (i.e., for biological-science majors, typically algebra-based) or university physics (i.e., for physics/engineering majors, typically calculus-based) (Tai & Sadler, 2001). In college physics courses, females received higher average grades than males, even when controlling for academic background and other variables (Tai & Sadler, 2001). In university physics courses, males and females receive similar average grades in introductory courses, however when demographics and academic background are controlled (e.g., high-school course grades, AP courses, and SAT math score), males receive higher grades than females, on average (Hazari, et al., 2007; Tai & Sadler, 2001).

These previous studies on students' grades did not account for the pedagogical practices in the university courses, although the effect of high-school pedagogy was reported to have statistically significant effects on university performance and some pedagogical practices had differential effects by gender (Hazari, et al., 2007). However, studies of male and female student grades in introductory courses using reformed, interactive teaching methods, such as PI and Cooperative Problem Solving, show similar results by gender. At the University of Colorado, Boulder, researchers report *no* statistically significant differences between males' and females' average final grades in

calculus-based introductory physics courses taught using interactive-engagement techniques, although they did not control for background differences (Kost, Pollock, & Finkelstein, 2007; Pollock, et al., 2007). At the University of Minnesota, in calculus-based university physics courses using Cooperative Problem Solving, males obtained slightly higher final grades ($p < 0.001$) than females, on average, although again differences in academic background were not taken into account (Docktor & Heller, 2008). In both of these samples, further inspection of exam scores revealed that males obtain higher exam grades than female students, on average (Docktor & Heller, 2008; Kost, et al., 2007). Researchers at the University of California, Davis described a change in the gender differences in final grades with the implementation of reformed, collaborative teaching methods in an algebra-based introductory physics course; however no summary statistics were reported (Potter, De Leone, Ishikawa, Blickenstaff, & Hession, 2001).

In summary, results are mixed on the effects of interactive teaching on reducing the gender gap in postsecondary introductory physics courses. While the results at Harvard University show that the gender gap is reduced and even eliminated on conceptual surveys through the use of interactive teaching methods, other studies show that the differences between males and females persist on these surveys even when using similar interactive pedagogies. Furthermore, these previous studies document some gender differences in final grades, even in interactive classrooms. However, the results on the impact of pedagogy would be strengthened through analysis with more sophisticated statistical techniques.

While a great deal of progress has been made in understanding gender differences in classrooms taught using PER methods, more research is needed to understand the effects of these pedagogies on gender differences in performance on conceptual tests and course achievement. By re-analyzing both previous data and additional data from introductory courses in mechanics and electricity and magnetism at Harvard University, we will have a better idea of the nature of the gender differences in introductory courses at this university, as well as add to the growing body of research on the effects of interactive teaching on gender gaps in performance on conceptual tests and course achievement.

2.2.3 Reformed teaching and differences in achievement, by race and ethnicity

Interactive teaching and racial differences in physics achievement

Although the PER community has made great strides to understand the impact of reformed, interactive teaching on the gender gap, much less work has been done to examine how these teaching methods impact the achievement of under-represented minority students. The few studies we found did not present statistics confirming the impact of different pedagogical strategies, but did present some evidence that interactive engagement methods may help retain under-represented minority students at college and in physics. For example, Beichner and Saul (2003) present data showing that failure rates are lower for students in courses taught using SCALE-UP and note that these reductions in failure rates are especially notable for females and racial- and ethnic-minority students. Other studies focus on special programs that target at-risk students. At Rutgers University, an alternative Extended Physics program is offered for science and

engineering students at risk of failure (Brahmia & Etkina, 2001a, 2001b). This course sequence emphasizes the social aspects of learning, by incorporating Investigative Science Learning Environments (Etkina & Van Heuvelen, 2007) and other mathematics interventions, and gives students more opportunities to interact with teaching staff. These interventions were linked with improvements in both achievement and retention (Brahmia, 2008).

Additional research has been conducted at majority-minority institutions. Outreach efforts at Florida International University, a Hispanic-serving public research university, focus on improving participation of traditionally under-represented minority and women students in physics. Research shows that FIU introductory physics courses that implement Modeling Instruction (Wells, Hestenes, & Swackhamer, 1995) in course sections lead to improvement in students' attitudes about physics and physics learning, as measured by the CLASS (Adams, et al., 2006; Brewe, Kramer, & O'Brien, 2009). Remarkably, these results are among the first to show students' attitudes shifting to become more expert-like after course in introductory physics. At Chicago State University, an urban, public university with a majority Black student population, research suggests that the traditional measurement tools of PER (e.g., conceptual inventories) do not capture the resources of this student population (Sabella, Coble, & Bowen, 2008). These authors present evidence that students are receptive to inquiry-style instruction and are willing participants in the scientific community. However, the standardized conceptual inventories, such as the FCI, do not appear to describe the formal knowledge that students demonstrate in class and interviews adequately (Sabella, et al., 2008).

While these studies provide invaluable insight into how reforms impact students at minority-serving institutions and programs, we still do not know much about the experiences of under-represented minority students in reformed, interactively-taught introductory physics courses in which they are the minority. There is a dearth of data on the performance, both on conceptual tests and course achievement, of under-represented racial- and ethnic-minority students as compared to their peers in traditional university settings.

What we can learn from other studies of under-represented racial minorities

Although there is limited data from the PER community on under-represented racial minorities, we can learn from research conducted in other communities. In a nationwide survey, Hazari, et al. (2007) found that Black and Hispanic students obtained lower average final grades in university introductory physics courses, even when controlling for differences in academic background. These results, however, do not take into account the teaching methods used in the university courses. Furthermore, physics is not the only STEM discipline to document the achievement gaps of Black or Hispanic students when compared to their White or Asian peers. Treisman's (1992) pioneering work in calculus courses in the 1970s revealed that 60% of Black students who completed a first-semester calculus course received grades of D or F. Furthermore, only two Black or Hispanic students received grades higher than a B in any calculus course. Further investigation of how students studied mathematics revealed dramatic differences between Black and Asian students. Black students often worked in isolation:

For the Black students it meant this: You wake up in the morning. You go to class. You take notes. You get your homework assignment. You go home. You do your homework religiously and hand in every assignment on time. You put in six

or eight hours a week of studying for a calculus course, just what the teacher says, and... you fail. (p. 385)

In contrast, observations of Chinese students revealed a very different pattern:

They studied calculus for about 14 hours a week. They would put in 8 to 10 hours working alone. In the evenings they would get together. They might make a meal together and then sit and eat or go over a homework assignment. They would check each others' answers and each others' English. (p. 385)

These observations by Treisman led to the development of a program that emphasizes group learning and community life, creating a “challenging, yet emotionally supportive academic environment,” (p. 368) including students of all ethnicities, but targeting under-represented minorities. The program not only created an intellectual community for students, but taught students *how* to work together, producing impressive results for both Black and Hispanic students. Participating Black and Hispanic students obtained grades similar to the better-prepared White and Asian students not in the program and many of them continued on to pursue careers in STEM disciplines (Treisman, 1992).

After Treisman's groundbreaking findings on the improvement in performance and retention of under-represented minority students in introductory courses, similar programs emerged at other institutions, such as the Meyerhoff Scholars Program at University of Maryland, Baltimore County (UMBC; Maton, Hrabowski, & Schmitt, 2000). Based on the research literature, Hrabowski and Maton (1995) list four sets of factors linked to academic success for African-American students: knowledge and skill development, support and motivation, monitoring and advising, and academic and social integration. The Meyerhoff Scholars at UMBC designed a scholarship program to address

these factors by creating a “family-like social- and academic-support system” for selected Black students². The program strives to create an academic community, encourage study groups, and give participants the opportunity to interact with faculty and staff through mentor programs, research internships, and personal counseling; each of these features are attributed to the program’s success (Maton, et al., 2000). Participating students achieved higher grade-point averages, graduated in STEM disciplines at higher rates, and were more likely to be accepted to graduate school than their peers (Maton, et al., 2000).

While these programs are large-scale university-wide interventions, they offer insight into pedagogical techniques that can be implemented at the individual course level to promote the achievement of under-represented minorities. Courses should emphasize collaboration among students and establish workshops in which students work together on problems (Bonsangue & Drew, 1995; Treisman, 1992). Nelson (1996) encourages the use of active-learning techniques, suggesting that “a failure to make effective use of these techniques is also (unintentionally) discriminatory against Blacks and other traditionally under-represented groups” (p. 172). Further studies suggest that the incorporation of minority students’ worldviews and prior knowledge can increase their course achievement and improve their attitudes toward science (cf. Lewis, 2009). However, just as in studies on the reduction of the gender gap, some studies suggest that these strategies help *all* students and therefore may not close the racial-achievement gaps (Lynch, Kuipers, Pyke, & Szesze, 2005).

Many of the reforms born out of physics education research, including Peer Instruction, incorporate features that have been shown to improve the academic

² After 1996 the program was opened up to all students, with a focus on under-represented minority students.

achievement of under-represented racial minorities in university-wide interventions. In this chapter, we take a first look at the impact of these teaching methods on Black and Hispanic student learning and achievement in introductory physics courses in which these students are in the minority.

2.2.4 Persistence and pedagogy

While scores on conceptual surveys and course performance are useful for helping understand how interactive teaching meets the more immediate goals of an introductory science course, we are also interested in the long-term effects. Specifically, do changes in pedagogy lead to greater student persistence in the sciences, specifically for females and under-represented racial and ethnic minorities? In Seymour and Hewitt's (1997) book, *Talking About Leaving: Why Undergraduates Leave the Sciences*, the authors categorize and rank the reasons that students abandon science majors at four-year colleges and universities, based on a three-year ethnographic study. In interviews and focus groups, the authors discovered that after loss of interest in science and interest in other majors, the third highest-ranked reason for leaving science was poor teaching. Over ninety percent of the students who switched out of science—and three-quarters of the students who remained in the science major—said they were concerned about the poor quality of teaching in their science courses. Specifically, students list the lack of faculty-student interaction, the 'coldness' of the classroom, lack of preparation and organization, and dullness of presentations as evidence of poor teaching. Students specifically call out the lecture method:

They just continually write. And they're standing in front of what they write, but they don't care. They'll look over their shoulder now and then, and say, 'Okay, you are all still there,' and just keep going. And the number of people that don't go to classes is amazing. (p. 154)

Students also compare the interactive nature of their non-science courses with the sterile nature of science lectures:

The classes in my new major seem to be about the same size, but there's so much more interaction between the professors and the students. In the math classes it seemed like the professor would just go up to the chalkboard and start doing problems... (p. 147)

Students describe how teaching could be improved in science courses, listing "openness, respect for students, encouragement of discussion, and the sense of discovering things together." (p. 148). Similarly, Strenta, et al. (1994) found that science majors at highly selective universities rated their instruction in science lower than did students with nonscience majors. Furthermore, although most students who switched out of a science major listed interest in another major as their primary reason for leaving, about 40% of students criticized the poor quality of instruction as cause for leaving science (Strenta, et al., 1994).

College student attrition from STEM majors most often occurs in the first or second year of college (Seymour & Hewitt, 1997). Furthermore, Manis, Sloat, Thomas, and Davis (1989; as cited in Seymour & Hewitt, 1997) found that students' experiences in their freshmen-science courses were the most influential in their decision to switch out of their major. In a study on retention in engineering majors, Lichtenstein, Loshbaugh, Claar, Bailey, and Sheppard (2007) found that poor teaching in pre-engineering courses could cause students to think that their engineering courses would be poorly taught and

therefore contemplated leaving the major. Encouragingly, these authors also found that “a single positive interaction, excitement about a course’s teaching and/or content... cause a student to confirm his or her choice to stick with engineering.” (p. 20). This finding offers promise for the impact of using interactive, engaging teaching methods in just a single introductory course to improve the retention of students in STEM majors. In the last section of this chapter, we examine whether changing the pedagogy in an introductory physics course can have a longer-term impact on student persistence in science and engineering. We also examine the effects by gender and race, documenting the effects of pedagogy on under-represented minority students’ persistence in the sciences.

2.2.5 Research questions

PI has been shown to greatly improve student learning, on average, as measured by conceptual surveys and exam scores. Previous research on gender and racial-minority issues in the classroom suggest that several key features of PI and other interactive teaching techniques offer promise at reducing gender and racial gaps in conceptual understanding, course achievement, and persistence. In this chapter, we examine the effect that exposure to PI has students by gender and race, by addressing the following research questions:

- (1) In an interactive classroom, what are the differences in student performance on tests of conceptual understanding among students of different genders and races?

- (2) In an interactive classroom, what are the differences in students' course achievement (*e.g.* final exams and final grades) among students of different genders and races?
- (3) What is the relationship between interactive teaching and retention in science majors? Furthermore, how does pedagogy affect the retention of females and racial minorities in the sciences?

2.3 Methods

2.3.1 Description of courses

In this study, we examine data from four different courses across 17 years. In Table 2.1, we display a timeline of the courses, including details of teaching methods used. Physics 1 courses were designed to provide a general introduction to physics for students majoring in the life sciences, satisfying pre-medical school requirements and requiring only limited use of calculus in the course. Physics 11 courses were intermediate-level introductory physics courses also designed for non-physics majors, but involving more advanced mathematics, requiring at least concurrent enrollment in multivariable calculus. Courses designated with an “a” at the end of their title are first semester in a two-semester sequence, covering topics in classical mechanics. Courses ending with “b” are second-semester courses, covering topics in electricity and magnetism, circuits, and optics.

Each course met for 1.5 hours, twice a week, in a large lecture hall. Smaller sections (15-20 students) were led by teaching assistants once a week for 1-2 hours. As summarized in Table 2.1, the teaching methods used in the courses from 1990-2007

differed over the years and are classified into three groups: traditional (T), partially interactive (IE1), and fully interactive (IE2). In 1990, both the lectures and section were taught in a traditional lecture format (T). In 1991, PI was introduced into the lectures, as described in the background section. Also at this time, reading quizzes were implemented to encourage students to read before class and, in 1995, the courses began using a research-based textbook developed by Mazur (in print). Through 1995, the sections remained largely traditionally taught, with a teaching assistant presenting to small groups of students seated in rows. As only parts of the courses from 1991-1995 were reformed, we refer to the teaching methods used as IE1.

In 1996, the traditional sections were reformed into workshops to engage students more actively. During the first hour of the workshops, students worked in small groups on *Tutorials in Introductory Physics* (McDermott, et al., 2002), giving them additional opportunities to discuss concepts with one another with structured hands-on activities. In 1997, additional time was added to the workshops to allow for more interactive group-work on the homework; teaching assistants led students through cooperative group problem-solving activities, similar to those developed by Heller, et al. (1992). Students again worked in small groups on problems from the textbook, while teaching assistants circulated to help students with difficulties and ask students to explain their work. Throughout the entire section, emphasis was placed on encouraging students to work together on solutions to problems, rather than obtaining the answer from the instructors. From 1996-2007, all courses were taught using these interactive teaching methods in lecture and section and are designated as IE2 in Table 2.1.

Table 2.1. Description and timeline of the courses in this study

Year	Course	Label	Teaching methods used		
			Pre-lecture	Lecture	Section
1990	11a	T	n/a	traditional	traditional
1991, 1993-5	11a	IE1	reading quizzes	PI	traditional
1996-7	11a	IE2	preclass reading	PI	<i>Tutorials</i> and, in 1997, some cooperative activities
1998, 2000	1a	IE2	JiTT	PI	<i>Tutorials</i> and cooperative problem-solving activities
2002-3, 2005-6	1b	IE2	JiTT	PI	<i>Tutorials</i> and cooperative problem-solving activities
2007	11b	IE2	JiTT	PI	<i>Tutorials</i> and cooperative problem-solving activities

In all years, the weekly homework consisted of traditional quantitative problems. Students were provided with a handout on problem-solving strategies from (Heller, et al., 1992) to encourage students to analyze the problem qualitatively, plan their solution, and check their answer to see if it makes sense. From 1991, the exams were reformed to incorporate conceptual problems, as well as traditional quantitative problems, to reinforce the importance of conceptual understanding in learning physics.

A total of five instructors taught these courses. One instructor (Mazur) taught all but one year³, in each of the courses (Physics 11a, 1a, 11b, 1b) and using each of the methods (T, IE1, IE2). Therefore, there was a great deal of consistency in the implementations of the interactive methods across years, although the findings may mask

³ In the one year (1997) in which Prof. Mazur did not teach, his postdoctoral research associate was one of the co-instructors, ensuring continuity in the use of the interactive teaching methods.

unobserved between-year and between-instructor differences that must be taken into account when interpreting the findings.

2.3.2 Measures

In this section, we describe the instruments we used in our analysis. We first describe the assessments of students' conceptual understanding—those administered at the start of the course are used as predictor variables, while those administered at the end of the course comprise part of our outcome variables. We also describe how we measured students' course achievement as well as how we determine student persistence in the sciences.

Conceptual understanding

In Physics 11a and Physics 1a, the FCI (Hestenes, et al., 1992) was administered to investigate differences in students' conceptual understanding of Newtonian mechanics. This test is used widely in the physics-education community to assess how instructional practice affects students' understanding of first-semester physics (e.g., Hake, 1998). The assessment was administered as a pretest, during the first week of classes, and again as a posttest, during the last week of class and leading up to their final exam. Students had 45 minutes to complete the tests on a computer, without the use of textbooks or other materials. Students received credit for participation in completing these assessments, but they did not receive extra credit for correct answers.

There were two versions of the FCI used in this study. From 1990-1994, the first version of the test (FCI-1) was administered, which contains 29 questions. Starting in 1995, the revised version of the test (FCI-2) containing 30 questions was used; 26 of

which are nearly identical to the original version. Consequently, our analysis of student achievement in the Physics 11a courses (1990-1997) uses only the 26 FCI questions common to both versions of the test. Student achievement in the Physics 1a courses (1998 and 2000) were analyzed using all 30 questions from FCI-2.

Unfortunately, the sample distributions of student scores on the posttest are highly skewed to the right, as many students obtained the maximum score on the test. While the fact that students are obtaining the maximum score on the test is encouraging for instruction, it introduces problems in linear regression analysis or other analyses that make the usual normal theory assumptions. Therefore, we collapsed students' posttest FCI scores into two broad categories, separating students with scores below and above 80%.

In Physics 11b and 1b, a pretest of student understanding of electricity and magnetism was not administered, based on the belief that such pretest scores would not have had much meaning as the terms and concepts in electricity and magnetism are often too unfamiliar to students who have not taken a relevant physics course. Instead, during the first two weeks of the courses, the FCI was administered to assess students' understanding of Newtonian mechanics, which is a building block for many concepts in electricity and magnetism. At the end of the semester, two weeks before the final exam, students' conceptual understanding of electricity and magnetism was measured using the CSEM (Maloney, et al., 2001), which contains 32 questions. Students had one hour to complete the assessments, without the use of the textbook or other materials. Students were again awarded credit for participation in completing these assessments. To better interpret the magnitude of the differences between males and females and students of

different races, we transformed students' scores into z-scores, subtracting the sample average score for each course from each individual student's score and dividing by the corresponding sample standard deviation. Therefore, the sample mean on the z-scores in each course is zero, while the sample standard deviation is one.

Course achievement

To investigate differences in course achievement by gender and race, we treated students' final exam scores and final grades as our outcome variables. The exams differed each semester, but within a given course we assumed that the tests were measuring similar constructs. Therefore, we combined exam scores across semesters within a given course by again transforming students' scores into z-scores. We repeated the process with final grades; all differences are reported in terms of the respective sample standard deviations.

We also looked in more detail at the components of students' course grades in Physics 1b. In this course, the final grade was computed on an absolute scale and was based on a point total of 100. Twenty points were awarded for the homework, 10 points for the reading assignments, 10 points for the laboratories, and 60 points for the examinations. Bonus points worth up to 5 points could be accumulated by responding to ConcepTests in class, although credit was given only for participation, not obtaining the correct answer. We examined students' grades in Physics 1b, by race, on each component.

Retention

To examine the relationship between exposure to interactive teaching and student retention in the sciences, students were asked at the beginning of Physics 11a courses

(1990-6) to indicate their major (we asked first-year students to indicate in which subject they planned to major.) We then compare their responses with their major recorded at graduation. Here, we focus on students who indicated that they were majoring in a physical, engineering, or biological science discipline at the start of their introductory physics course, examining which students switched from a science to a nonscience major.

2.3.3 Sample

Over 2400 students were enrolled in the courses that provided the achievement data for the analyses in this chapter; here we describe these students briefly and identify which students completed the assessments under study. In Table 2.2, we present univariate descriptive statistics for the students in Physics 11a (1990-1, 1993-7), Physics 1a (1998, 2000), Physics 11b (2007), and Physics 1b (2002-3, 2005-6). We list statistics on student demographics, including the total number of students, percentage of male students, and racial breakdown of students enrolled in the course, and students' majors.⁴ In Physics 11a and 11b more males than females were enrolled; however for Physics 1a and 1b, more females were enrolled.

Most of the students in these courses were Asian-American or White. Only about 6% of the students in the Physics 11 courses and 12% in the Physics 1 courses were Black, as reported by the Office of the Registrar. We also document similarly low percentages of Hispanic students. In our preliminary analyses, we considered students of each race separately and found that not only did the results for Black and Hispanic students differ from the rest of the students in the course, but that these two groups

⁴ Note that at the time of data collection (January 2009), some students in Physics 11b and 1b had not yet graduated. The data we present for these students is their declared major at the time of data retrieval from the Office of the Registrar.

obtained similar results, on average. Also considering that students of both races are under-represented in the scientific disciplines, we combined students of these races into one group, which we then refer to as under-represented minorities (URM). In our analyses, we compare the academic performance of these under-represented racial- and ethnic-minority students with that of the rest of the students in the courses.

Table 2.2. Univariate descriptive statistics on selected student demographics and majors in all four courses in our study.

	Course			
	11a	1a	11b	1b
<i>N</i>	1411	325	130	537
male	64%	45%	54%	34%
Asian-American	35%	27%	26%	22%
Black	6%	12%	6%	11%
Hispanic	5%	9%	7%	9%
Foreign	1%	3%	15%	7%
Native American	0%	1%	2%	1%
White	36%	34%	31%	41%
Other/unknown	18%	14%	13%	9%
science major	89%	39%	93%	52%
nonscience major	11%	61%	7%	48%

Throughout our analysis, we removed students from the analysis who did not complete the final exam. Furthermore, to deal with outliers, we removed from our analyses all students with final grades more than three standard deviations lower than the mean for their course. This resulted in a removal of 23 students from 11a, 2 students from 1a, 1 student from 11b, and 6 students from 1b.

Although students were awarded credit for participation on the conceptual tests, not all students completed these assessments. Approximately 20% of the students in our

sample did not take either the pretest or posttest conceptual test in their course. In Table 2.3, we compare the scores of students with no missing (matched) datasets in Physics 1b with the scores of those students with missing data. The results in this course resemble the data in the other courses.

Table 2.3. Comparison of Physics 1b students with no missing data and students who did not complete either the conceptual pretest nor the posttest in Physics 11a on student demographics, FCI posttest performance, final exam z-scores, and final grade z-scores.

	no missing data			missing data			difference
	<i>N</i>	mean	SD	<i>N</i>	mean	SD	
% male	471	0.32	0.47	99	0.36	0.48	-0.04
% URM	471	0.19	0.29	99	0.22	0.42	-0.03
CSEM z	471	0.01	0.99	17	0.00	1.08	0.01
final exam z	471	0.08	0.98	99	-0.26	1.03	0.34
grade z	471	0.15	0.94	99	-0.43	0.98	0.58

We note that slightly more males than females and slightly more under-represented minority students are missing data. However, there were not large differences in the average performance on the CSEM posttest between these two groups. However, the differences in course achievement between students with no missing data and those with missing data are statistically significant ($p < 0.001$). Students with no missing data obtained final exam scores that were 0.34 standard deviations higher than students with missing data, on average. The difference between these two groups of students is even higher on final grade—0.58 standard deviations.

While in later chapters we use statistical techniques intended to resolve issues of data-missingness, we include only those students that had complete data in this analysis, for two reasons. First, we would like to present our results in light of other published

work (e.g., Kost, et al., 2009), which only used students with complete datasets. Second, we are most interested in generalizing to the population of students who will put the effort forth to complete the course assignments. Our focus in this chapter is on gender and racial differences in average learning and achievement. By narrowing our lens to exclude those students who did not complete these assignments, we remove students who perhaps did not take the course seriously, but we recognize at the same time that these results do not represent the entire population of students in these courses.

2.3.4 Analytic methods

To investigate differences between male and female students and different racial groups on tests of conceptual understanding and course achievement, we used multiple regression analysis to examine the relationship between our dependent variable, e.g., CSEM posttest scores or final grades, and selected independent variables, e.g., MALE or FCI pretest scores, at once. In each of our analyses, we first present fitted regression models⁵ which only include predictors describing gender and race, using the dichotomous variables MALE (1 = male, 0 = female) and URM (1 = under-represented minorities, 0 = everyone else). We then control for variables describing the students' background preparation, adding FCI pretest and SAT scores to the statistical models, and then re-estimating the slope parameters for MALE and URM. We perform these analyses separately for each course, as the goals and assessments are different in each.

⁵ In each regression model, we fit data to a given model (equation) to estimate the intercept and slope parameters. For example, if we want to predict CSEM scores, one regression model could do so using only gender: $CSEM = \beta_0 + \beta_1 MALE$. Another regression model could fit an equation with both gender and race: $CSEM = \beta_0 + \beta_1 MALE + \beta_2 URM$. Regression analysis determines the best intercept and slope parameters to fit our data.

As described in the *Measures* section, we do not use students' raw FCI posttest scores as our outcome variable in mechanics courses; the sample distribution of these scores was highly non-normal, as many students obtained the maximum score on the survey. Instead, we collapsed students' scores into two categories: FCIhigh (posttest scores $\geq 80\%$) and FCIlow (posttest scores $< 80\%$). Therefore, to examine gender and racial differences on the FCI, we model whether or not a student scored high on the posttest as a function of our independent variables MALE and URM. We cannot use linear regression, because our outcome variable can only take on the values of 0 or 1. Instead, we use logistic regression analysis, which represents the probability of switching out of a science major as a logistic function of predictors x , as follows:

$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

The population odds of scoring high (versus low) can be obtained by dividing by one minus the probability: $(odds = \frac{P}{(1-P)})$. In logistic regression analysis, the log-odds of scoring high has a linear relationship with the hypothesized independent variables. Maximum likelihood estimation⁶ is then used to estimate the intercept (β_0) and slope parameters by fitting the following model to the data:

$$\text{Log - odds}(\text{scoring high}) = \beta_0 + \beta_1 \text{MALE} + \beta_2 \text{URM} + \beta_3 \text{FCIpre} + \beta_4 \text{SATmath}$$

⁶ The best fits are determined using maximum likelihood estimation, which selects the parameter estimates that maximize the likelihood of observing the data given the model. See (Hosmer & Lemeshow, 2000) for additional details on logistic regression.

Algebraic manipulation results in the following statistical model, specifying the hypothesized relationship between the population probability of scoring ‘high’ and the independent variables.

$$P(\text{of scoring high}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \text{MALE} + \beta_2 \text{URM} + \beta_3 \text{FCIpre} + \beta_4 \text{SATmath})}}$$

To compare the performance of different groups of students we do not focus on the intercept estimate, but instead interpret the parameter estimates associated with the predictors in terms of fitted probabilities, odds and odds-ratios.

In our findings, we present the results of fitted several logistic regression models. In our first models, we only include the variables MALE and URM in order to describe the total effect of the gender and race differences on the outcome, before controlling for differences in background preparation. We then add independent variables describing the FCI pretest and SAT scores to control for individual differences on these assessments prior to the start of the course. To compare males’ and females’ performances, we estimate their fitted odds of scoring high by presenting the fitted odds ratios we obtain from the regression equation, which can be obtained by manipulating the log odds equation above:

$$\frac{\text{odds (for a male to score high)}}{\text{odds (for a female to score high)}} = e^{\beta_1}$$

For analyses of the Physics 11a data, we also include and estimate a parameter to represent the effect of pedagogy: IE2. This slope parameter variable captures the impact of the fully interactive methods used in fall 1996 and 1997. To investigate whether the presence of the particular teaching method has a differential effect by gender, we included a two-way interaction between gender and pedagogy (MALE \times IE2) as an additional predictor in the Physics 11a regression analyses. Basically, the inclusion of this interaction permits the gender gap to differ between IE1 courses and IE2 courses. The slope parameter of this interaction variable represents the differential impact of IE2 courses as compared to IE1 courses on the gender gap.

To examine the effects of gender and race on students' performance on tests of conceptual understanding in the electricity and magnetism courses, we use linear regression analysis to predict CSEM posttest scores. Again, we fit several models to examine gender and race differences. In our first model, we include the principal question predictors MALE and URM in order to estimate the total effect of these variables; then, we then control for the prior measures FCI and SAT scores in subsequent models. We estimate differences between males and females, and students of different races, on the CSEM from the fitted models and present them in units of standard deviations.

For data on all courses, we use linear regression analysis to examine the effects of gender and race on course achievement, focusing on students' performances on the final exam and their final grades. Again, we first examine the total effects of these predictors before controlling for background variables by adding the FCI and SAT scores to our models. For a more in-depth look at course grades by race, we compare non-URM and

URM student grades on the myriad components of their grades in Physics 1b, using the Wilcoxon rank-sum test⁷ to compare medians between groups.

To understand the effects of pedagogy on student retention, we use multilevel logistic regression analysis to predict the odds that a student will switch from a science to a nonscience major after taking Physics 11a. We first fit logistic regression models to control for students' background variables (e.g., their year in school and background preparation) then add the main effect of pedagogy. The associated parameter indicates whether the teaching methods used in an introductory physics have any effect on student retention in a science major. Finally, we control for course achievement, using final grades and FCI posttest scores, in order to examine how student performance, achievement, and pedagogy all relate to student retention simultaneously.

2.4 Results

2.4.1 Effects of gender and race on students' conceptual understanding

In this section, we present results on students' performance on conceptual surveys. Because we have data on their performance in four different courses, we present the overall findings at the start of each subsection, before delving into the details.

Results from the Force Concept Inventory.

Before we examine the specifics of the multiple regression analysis, we provide a short summary of our findings to help draw the reader's attention to our main points, as follows:

⁷ This statistical test is non-parametric, which means that it does not make the assumption that the scores are drawn from a normal (Gaussian) distribution.

- In the more advanced introductory course (Physics 11a), students who were taught using IE2 methods outperform students taught using IE1 methods, on average;
- In Physics 11a, the average gender “gap” is larger in IE1 than IE2 courses, even after we control for differences in student background;
- There are no statistically significant gender differences in outcomes in Physics 11a courses taught using IE2 methods, even before controlling for differences in student background;
- When we examine outcomes from the lower-level introductory course (Physics 1a), gender differences remain statistically significant even when we account for differences in student background;
- When investigating differences by race, in both Physics 11a IE1 and IE2 courses, we find a “gap” in performance between the races, which is only statistically significant once we control for differences in student background;
- However, in Physics 1a, the racial “gap” remains statistically significant even after we control for differences in student background.

In Table 2.4, we present fitted multilevel logistic-regression models in which we predict the odds that a student will score ‘high’ (above 80%) on the FCI posttest in introductory mechanics courses (versus “low”). In the table, we present the parameter (intercept and slope) estimates with associated p -values, as well as an overall goodness-

Table 2.4. Fitted multilevel logistic-regression models predicting whether a student scores high on the FCI, controlling for student demographics and background, in (a) Physics 11a IE1 and IE2 courses ($N = 993$) and (c) Physics 1a IE2 courses ($N = 305$). Each column represents a different model to our data, using the variables in the left-most column.

(a)

	Model 1	Model 2	Model 3	Model 4	Model 5
intercept	1.14***	0.88**	0.80***	-4.44***	-10.57***
male	1.49***	1.49***	1.69***	0.95***	0.94***
URM	-0.71*	-0.73**	-0.76**	-0.36	-0.05
IE2		0.85**	1.31**	1.71***	1.58**
malexIE2			-1.09*	-1.18*	-1.13*
FCI pretest				0.86***	0.81***
SAT math					0.86**
-2LL	765.2	759.8	754.7	585.0	572.4

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(b)

	Model 1	Model 2	Model 3
intercept	-0.16	-2.52***	-4.51*
male	1.32***	0.71*	0.71*
URM	-0.99**	-0.85*	-0.72*
FCI pretest		0.64***	0.66***
SAT math			0.25
-2LL	366.9	315.4	305.2

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

of-fit statistic, negative two times the log likelihood $(-2LL^8)$; with which lower values indicate a better fit of the hypothesized model to the data). Note that we do not include data from the 1990 traditional course, as the FCI pretest was not administered that year. The estimates of the intercept and of the slope parameters associated with each predictor do not have readily accessible interpretations. Instead, to address our research questions, we transform these estimates to obtain the estimated odds of scoring high (versus low), by gender and race. We compare the odds for different groups of interest by looking at the fitted odds ratios, e.g., for males and females; URM and nonURM; or IE1 and IE2 courses.

In Table 2.4a, we show the fitted logistic regression models for Physics 11a courses. Each model shows the best fit for a given equation to the data. For example, the fitted values of the outcome for Model 1 are given by the equation:

$$\hat{P}(\text{scoring high on } FCI\text{post}) = \frac{1}{1 + e^{-(1.14+1.49*MALE-0.71*URM)}}$$

The estimates for the slope parameters associated with question predictors MALE and URM indicate that, for this population, both race and gender are statistically significant predictors of the probability that a student will score high on the FCI. The equation can be manipulated to obtain the fitted odds of scoring high:

$$\widehat{odds}(\text{scoring high on } FCI\text{post}) = e^{(1.14+1.49*MALE-0.71*URM)}$$

⁸ As mentioned in the methods section, logistic regression analyses uses maximum-likelihood estimation (as opposed to least-squares) to determine the parameter estimates that best fit to our data. In this fitting procedure, the likelihood function is maximized, but for ease of computation, it is actually the log of the likelihood function that is actually maximized (and the negative log likelihood minimized). We use -2 times the obtained log likelihood as a goodness-of-fit statistic because it has a chi-square distribution in large samples and therefore differences in its value between models can be used to test the null hypothesis that the slope parameters of our predictors are zero, in the population.

We find that the fitted odds that a male student will score high are about 4.5 times the fitted odds that a female student will do the same, on average. For non-URM student, the corresponding fitted odds are twice those of URM students.

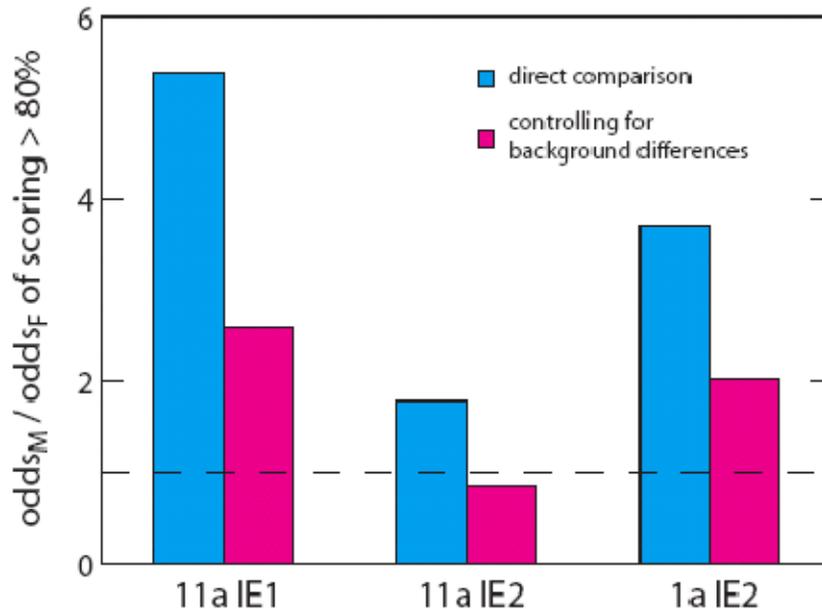
In Model 2, we add the effect of pedagogy, IE2, as a predictor to distinguish the effects between partially-interactive and fully-interactive teaching methods. The slope estimates indicates that the odds that a student in the IE2 courses will score high are 2.3 times the corresponding odds for a student in the IE1 courses, on average. When we include an interaction between pedagogy and gender as a predictor in the model, $\text{male} \times \text{IE}$, we find that the impact of pedagogy differs for male and female students (Model 3; $p < 0.01$). In an IE1 course, the fitted odds that a male student will obtain a high score on the FCI are more than five times the fitted odds for female students. In contrast, in an IE2 course, the fitted odds for males are less than twice the fitted odds for females and this difference is not statistically significant ($p = 0.17$).⁹ These results indicate that there is a “gender gap” in FCI performance in IE1 courses, but not in IE2 courses.

In Models 4 and 5, we now control for students’ FCI pretest and their scores on the SAT math section, both of which are statistically significant predictors of whether or not a student scores high on the FCI. We find that the higher a student’s pretest or SAT math score, the higher the odds that the student score high on the FCI posttest. Controlling for these background variables, however, race is no longer a statistically significant predictor of high FCI scores, however the impact of gender and pedagogy remain statistically significant. Furthermore, the interaction between pedagogy and gender is still a statistically significant predictor of the outcome when academic

⁹ We obtained the p -value from a post-estimation test in Model 3.

background differences are controlled. In IE1 courses, the fitted odds that a male student will score high are still more than 2.5 times the odds for female students with the same FCI pretest and SAT math scores. In IE2 courses, however, the fitted odds that a female student will score high on the FCI posttest are now *greater* than male students' odds of scoring high ($\text{odds}_{\text{male}} = 0.83 \times \text{odds}_{\text{female}}$) although postestimation tests on Models 4 and 5 show that gender differences in IE2 courses are not statistically significant ($p = 0.7$). In Figure 2.6, we present and compare these fitted odds-ratios for males and female students in IE1 and IE2 Physics 11a courses.

Figure 2.6. Fitted odds of scoring high (> 80%) on the FCI posttest in Physics 11a IE1 and IE2 courses, as well as Physics 1a, by gender. Dashed line indicates even odds ($\text{odds}_{\text{M}} / \text{odds}_{\text{F}} = 1$). The blue bar presents the fitted odds-ratios before controlling for background differences (Model 1), the pink bar shows fitted odds-ratios after controlling for FCI and SAT math scores (from Model 5 in Table 2.4a, Model 3 in Table 2.4b).



We present similar fitted logistic regression models for students in Physics 1a in Table 2.4b. In the lower-level introductory mechanics course, we also find statistically significant differences in the fitted odds of scoring high for students of different genders and races. On average, in this population, the fitted odds that a male student will score high on the FCI posttest are almost four times the fitted odds for female students. Under-represented minority students obtain, on average, fitted odds of scoring high that are about 1/3 of the fitted odds obtained by their non-minority peers.

When we control for students' performance on the FCI pretest in Model 2, unlike our results for students in Physics 11a, race and gender remain statistically significant predictors of the odds of scoring high on the FCI for students in Physics 1a. On average, the fitted odds of scoring high on the FCI for under-represented minority students are less than half the fitted odds for their non-minority peers with the same FCI pretest score. The fitted odds for male students are twice the fitted odds for female students with the same FCI pretest score. SAT math (and verbal, not shown) does not have a statistically significant effect in the final model (Model 3).

In Figure 2.6, we present and compare the fitted odds-ratios estimated in Physics 1a with those obtained in Physics 11a, for male and female students. These fitted odds-ratios compare males' and females' odds of scoring high on the FCI posttest. We obtain the largest ratios in the Physics 11a IE1 courses, even after controlling for background differences. In the Physics 11a IE2 course, the fitted odds-ratios are smaller and when we control for background differences, the fitted odds that a male will score high are not statistically significantly different than the fitted odds that a female will score high. Although Physics 1a is taught using these IE2 methods, the odds-ratios are again

comparable to the upper-level introductory course taught using IE1 methods. When background differences are controlled, the fitted odds that a male will score high are twice the fitted odds for a female.

In summary, we re-analyzed the FCI data from Physics 11a and 1a, examining the effect of gender on students' posttest performance. We first find that in the more advanced introductory courses, there is a statistically significant effect of pedagogy on the gender gap. In 11a courses that use the IE2 teaching methods, the disparity between males' and females' fitted odds of scoring high on the FCI is smaller than the courses taught using IE1 methods. Furthermore, the fitted odds were not different for males and females in IE2 courses. However, once we examine the lower-level introductory course (Physics 1a), the gender gap reappeared even though the course was taught using IE2 methods

We also found differences across courses when examining the impact of race on FCI posttest performance. In Physics 11a, race was not a significant predictor of scoring high on the FCI once we controlled for background differences. However, we found that there were still statistically significant differences in Physics 1a even after controlling for background differences: the fitted odds of scoring high on the FCI for nonURM students are still about twice those of URM students.

Results from the Conceptual Survey on Electricity and Magnetism

In this section, we investigate the relationships between gender/race and students' conceptual understanding in introductory physics courses, covering topics in electricity and magnetism; all courses were taught using IE2 methods. We again presage our main findings briefly:

- We find statistically significant gender differences on the CSEM in both Physics 11b and Physics 1b, on average. When we control for pretest scores, however, these differences are no longer statistically significant or become marginally statistically significant.
- In the more advanced introductory course (Physics 11b), we find no statistically significant racial differences, on average. In Physics 1b, these differences are present before we control for differences on the pretest, but no longer statistically significant once we include this variable.

We now present the details of our findings, grouped by course. In Table 2.5 we show fitted linear regression models in which Physics 11b students' scores on the CSEM posttest (transformed to z-scores) are predicted. We present both the sum-of-squared errors (SSE) and R^2 statistics to summarize how well these hypothesized models fit the data. Again, each model shows the best fit for a given equation to the data. For example, the fitted values of the outcome for Model 2 are given by the equation

$$\widehat{CSEM} = -2.88 + 0.02 * MALE - 0.05 * URM + 0.35 * FCI$$

The slope estimates in the table provide the estimated differences between male and female students (as well as between URM and nonURM students) in units of the standard deviation of the CSEM scores. The slope associated with FCI represents the estimated difference in CSEM scores per 10% difference in FCI pretest scores, again in units of standard deviations.

In Model 1, we find that gender predicts students' CSEM posttest scores ($p < .05$); males obtained CSEM scores about 0.37 standard deviations higher than did female students, on average. The effect of race, however, was not statistically significant in Physics 11b. However, when we control for FCI scores, we find that gender is no longer a statistically significant predictor of CSEM scores.

Table 2.5. Fitted linear regression models predicting Physics 11b CSEM z-scores using selected background and demographic variables ($N = 91$, one outlier removed).

	Model 1	Model 2
intercept	-0.20	-2.88****
male	0.37*	0.02
URM	-0.25	-0.05
FCI		0.35***
SSE	71.0	45.1
R ²	0.05	0.40

* $p < 0.05$, *** $p < 0.0001$

In Table 2.6, we present linear regression models predicting Physics 1b students' CSEM posttest scores. In Model 1, we find that, on average, males outscored females by 0.42 standard deviations. We also find a statistically significant effect of race in predicting CSEM scores. URM students obtained average scores that were approximately 0.39 standard deviations lower than their peers. When we control for FCI scores in Model 2, the effect of gender is still marginally statistically significant. Male students obtained CSEM posttest scores that were 0.15 standard deviations higher than their female peers

with the same FCI scores, on average. However, we find that when we control for this background knowledge,¹⁰ the effect of race is no longer statistically significant.

When looking at all of the results on assessments of students' conceptual understanding, there are some key trends revealed. In terms of gender differences, we find that males outperform females, on average, except in the advanced introductory-

Table 2.6. Fitted linear regression models predicting Physics 1b CSEM z-scores using selected background and demographic variables ($N = 471$).

	Model 1	Model 2
intercept	-0.06	-1.93***
male	0.42***	0.15~
URM	-0.39***	-0.08
FCI pretest		0.27***
SSE	426.1	317.8
R ²	0.07	0.30

*** $p < 0.0001$

level mechanics course (Physics 11a) taught using IE2 methods. When we control for background differences, the gender differences largely disappear, except in the lower-level mechanics course (Physics 1a). IE2 pedagogies appear to help all students in the advanced course 'max' out the test, but our results from the lower-level introductory mechanics course mirror the results from other institutions. We also find that URM students consistently obtain lower average scores on the conceptual assessments, but these differences are largely accounted for by background differences in our sample. In the next section, we move away from examining students' performance on low-stakes

¹⁰ The effects of SAT math and verbal scores were not statistically significant in predicting CSEM scores in Physics 11b and Physics 1b and are therefore not shown in these regression models.

conceptual assessment and investigate the effects of gender, race, and pedagogy on students' achievement.

2.4.2 Effects of gender and race on students' course achievement

Results from final exams

In this section, we present data on students' final exams in Physics 11a, 1a, and 1b. Remarkably, we find consistent findings across these courses in terms of both gender and racial differences:

- Consistently, males obtain higher average final exam grades than females, although once we control for background, these differences are no longer statistically significant;
- In contrast, URM students obtain lower average final exam grades consistently and these racial differences remain statistically significant when we control for selected background-preparation variables.

In Table 2.7, we show fitted regression models predicting students' exam z-scores on final exams in introductory mechanics courses. Each model shows the best fit for a given equation to the data. For example, the fitted values of the outcome for Model 4 in Table 2.7a are given by the equation:

$$\widehat{final\ exam} = -4.95 + 0.03 * MALE - 0.41 * URM + 0.21 * FCIPretest + 0.34 * SATmath + 0.14 * SATverbal$$

Table 2.7. Fitted linear regression models predicting students' final exam z-scores by student demographics and background variables in (a) Physics 11a IE1 and IE2 courses ($N = 993$) and (c) Physics 1a IE2 courses ($N = 305$). Note that each model represents the fitting of a different to our data, with the variables in the left-most column as predictors.

(a)

	Model 1	Model 2	Model 3	Model 4
intercept	-0.04	-0.06	-1.63***	-4.95***
male	0.32***	0.36***	0.00	0.03
URM	-0.76***	-0.76***	-0.58***	-0.41***
IE2		0.08		
malexIE2		-0.16		
FCI pretest			0.28***	0.25***
SAT math				0.34***
SAT verbal				0.14***
-2LL	785.7	784.5	629.8	588.9
ρ	0.08	0.09	0.27	0.31

*** $p < 0.001$

(b)

	Model 1	Model 2	Model 3
intercept	-0.01	-0.82***	-4.93***
male	0.37**	0.08	0.10
URM	-0.68***	-0.56***	-0.37**
FCI pretest		0.20***	0.14***
SAT math			0.42***
SAT verbal			0.17~
SSE	259.0	224.4	205.2
R^2	0.11	0.24	0.30

** $p < 0.01$, *** $p < 0.001$

In Table 2.7a, we present regression models fitted to data on students enrolled in Physics 11a courses. For Model 1, we present parameter estimates describing the effect of gender and race, without controlling for differences in background. We find that, on average, male students obtain final exam scores that are about one-third of a standard deviation higher than female students. Furthermore, URM students obtain lower average final exam scores than their peers, by more than three-quarters of a standard deviation. In Model 2, we investigate the impact of pedagogy on gender differences in the exam z-scores. Notice that the interaction variable, *malexIE*, is not statistically significant, indicating that the gender differences do not differ based on the teaching methods used in the course.

In Model 3, we control for background preparation in mechanics by including the FCI pretest scores and we find that the differences between males and females are no longer statistically significant. However, URM students obtain scores that are more than half of a standard deviation lower than their peers, when controlling for FCI pretest scores, on average. This difference remains statistically significant when we also control for SAT math scores; URM students score more than a third of a standard deviation lower than their peers with the same FCI pretest and SAT math scores, on average ($p < 0.001$).

We present fitted regression models for data on students in Physics 1a, in Table 2.7b. Again, both gender and race have statistically significant impact on students' final exam scores, as seen in Model 1. On average, in the population, male students obtain final exam scores over one-third of a standard deviation higher than female students ($p < 0.01$). Furthermore, the average difference between URM students and their peers is

similar to the gap in Physics 11a, about two-thirds of a standard deviation ($p < 0.001$), on average. When we control for FCI pretest scores in Model 2, we find that the average difference in outcome between the genders is no longer statistically significant. However, URM students' average scores were still more than a half of a standard lower than the average scores of their peers with the same FCI pretest score. In Model 3, we also control for SAT math scores, and find that the effect of race on the outcome is still statistically significant in Physics 1a. On average, in the population, URM students obtain final exam scores 0.37 standard deviations lower than their peers with the same FCI and SAT math scores ($p < 0.01$).

We do not present findings for final exam scores in Physics 11b, as we did not find either gender or race differences in our sample of students with no missing data, even though there are statistically significant differences in final exam scores between under-represented minority students and the rest of the class when all students in the course are included. In Table 2.8, we present fitted linear regression models in which we predict the for Physics 1b final exam z-scores, to examine the effects of race and gender on course achievement in the lower-level introductory course in electricity and magnetism. In Model 1, we find that the effect of gender on students' final exam scores is statistically significant; male students obtain final exam scores about two-tenths of a standard deviations higher than female students, on average ($p < 0.05$). In Model 2, we show that these gender differences are no longer statistically significant when FCI pretest scores are controlled.

We again find strongly statistically significant differences between students of different races in Physics 1b. On average, in the population, URM students obtain final

exam scores that are 0.58 standard deviations lower than their peers ($p < 0.001$). In Models 3 and 4, we control for differences in background preparation using FCI pretest scores and SAT scores, however there is still a marginally statistically significant effect of race on final exam scores ($p = 0.06$). On average, we predict that URM students obtain a final exam score 0.20 standard deviations lower than their peers with the same FCI pretest and SAT math and verbal scores.

Table 2.8. Fitted linear regression models predicting Physics 1b final exam z-scores by selected background and demographic variables ($N = 471$).

	Model 1	Model 2	Model 3
intercept	0.13*	-1.48***	-4.07***
male	0.20*	-0.03	
URM	-0.58***	-0.32**	-0.20~
FCI pretest		0.23***	0.20***
SAT math			0.20*
SAT verbal			0.17*
SSE	401.5	322.2	311.7
R ²	0.07	0.25	0.27

~ $p < 0.10$, * $p < 0.05$, *** $p < 0.001$

Overall, then, we find that males obtain higher final exam scores than females in these courses, on average. When we look at students with similar academic background scores, however, these differences are no longer statistically significant. Surprisingly, this is not the case for URM and nonURM students.

Results from final grades

In this section, we investigate gender and race differences in student grades in all of the introductory courses. We first briefly present our main findings:

- In the mechanics courses (Physics 11a and 1a), males obtained slightly higher average grades than females, but again once we control for selected background-preparation variables this difference is no longer statistically significant;
- In the electricity and magnetism courses (Physics 11b and 1b), there were no statistically significant gender differences (although females slightly outperformed males, on average, when we controlled for selected background variables);
- Finally, the story remains consistent for URM students: they obtain lower average scores as compared to their peers on final grades, even when pretest and SAT scores are accounted for.

We examine students' final grades as a measure of their achievement in introductory physics. In Table 2.9 we show fitted linear regression models of students' final grades in introductory mechanics courses, focusing on the effects of gender and race. Again each model represents a different equation with which we fit our data, an example of which we show below (Model 1 from 2.9b).

$$\widehat{final\ grade} = 0.20 + 0.19 * MALE - 0.46 * URM$$

We present the results for students in Physics 11a courses in Table 2.9a and Physics 1b in Table 2.9b. From Model 1, we find that both gender and race have statistically significant effects on students' final grades. On average, male students obtain final grades that are about a quarter of a standard deviation higher than female students (p

< 0.001). We find, in Model 2, that the difference between male and female students is no longer statistically significant when we control for FCI scores. The average final grade for URM students is estimated to be 0.74 standard deviations lower than their peers ($p < 0.001$). When we inspect Models 2 and 3, we find that we do not account for all of this difference by controlling for students' background preparation. On average, URM students obtain final grades that are 0.38 standard deviations lower than their non-URM peers with the same FCI and SAT math score ($p < 0.001$).

We present similar fitted regression models for students in introductory electricity and magnetism courses, in Tables 2.10 and 2.11. We find that the gender differences in Physics 11b and Physics 1b are not statistically significant in Model 1 in Tables 2.10 and 2.11. However, we detect statistically significant average racial differences in both courses, similar to the differences we observed in introductory mechanics courses. In both Physics 11b and Physics 1b, URM students obtain average final grades that are over half of a standard deviation lower than their peers before controlling for differences in background preparation ($p < 0.05$). This gap is reduced to about one-third of a standard deviation after controlling for FCI pretest and SAT math scores, as seen from Model 3 in Tables 2.10 and 2.11.

From these fitted regression models, we observe racial differences on measures of course achievement consistently, but not on the conceptual surveys, when controlling for background differences. To further examine this trend, we display the estimated average racial differences in Physics 1b on measures of conceptual understanding and course achievement in Figure 2.7. On the CSEM, we find statistically significant differences between URM students and their peers, although when we control for differences in their

Table 2.9. Fitted linear regression models predicting students' final grades z-scores by selected student demographics and background variables in (a) Physics 11a IE1 and IE2 courses ($N = 993$) and (b) Physics 1a IE2 courses ($N = 305$). Note that each model represents fitting a different model fitted to our data, including the predictors in the left-most column.

(a)

	Model 1	Model 2	Model 3	Model 4
intercept	0.04	0.00	-1.31***	-5.09***
male	0.27***	0.32***	-0.00	0.03
URM	-0.74***	-0.75***	-0.59***	-0.40***
IE2		0.14		
malexIE2		-0.18		
FCI pretest			0.24***	0.20***
SAT math				0.41***
SAT verbal				0.14***
SSE	620.8	619.2	507.1	460.9
R ²	0.09	0.09	0.26	0.32

*** $p < 0.001$

(b)

	Model 1	Model 2	Model 3
intercept	0.21	-0.27**	-3.14***
male	0.20**	0.03	0.02
URM	-0.46***	-0.39***	-0.24**
FCI pretest		0.11***	0.07***
SAT math			0.41***
SSE	115.9	104.5	93.1
R ²	0.11	0.19	0.28

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.10. Fitted regression models predicting Physics 11b final grade z-scores by selected background and demographic variables ($N = 93$).

	Model 1	Model 2	Model 3
intercept	0.18	0.26**	-3.90***
male	0.10	-0.17	
URM	-0.55*	-0.40*	-0.34~
FCI		0.26***	0.23***
SAT math			0.30~
SSE	52.2	37.7	36.5
R ²	0.01	0.33	0.35

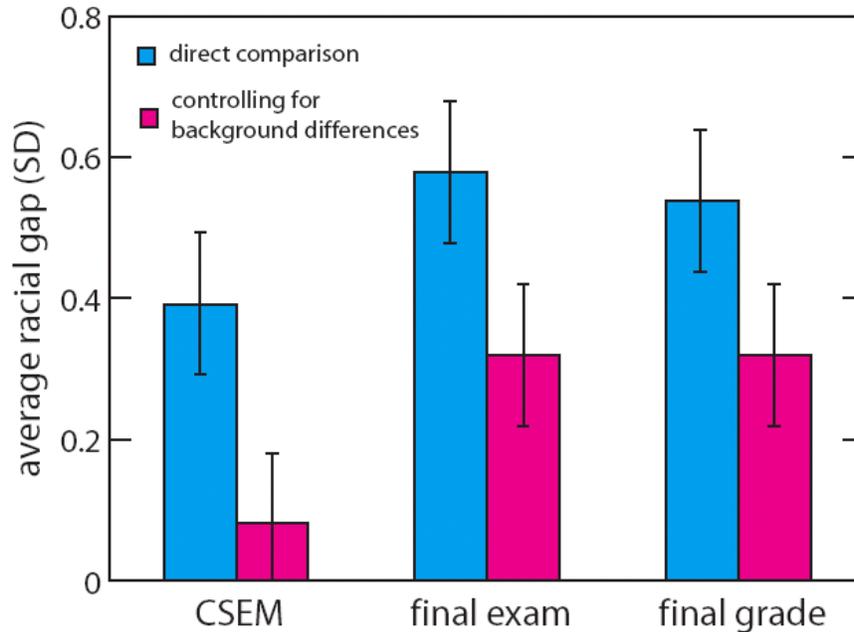
~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.11. Fitted linear regression models predicting Physics 1b final grade z-scores by selected background and demographic variables ($N = 471$).

	Model 1	Model 2	Model 3
intercept	0.24	-1.09***	-2.46***
male	0.03	-0.16*	-0.17*
URM	-0.54***	-0.32***	-0.26**
FCI		0.19***	0.18***
SATmath			0.20**
SSE	315.0	260.9	256.4
R ²	0.06	0.22	0.23

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 2.7. Estimates for the average differences between under-represented minority students and their peers in Physics 1b. The bars represent the magnitude of the gap, in units of standard deviation.



FCI pretest scores, the gap is no longer statistically significant. On measures of course achievement, however, the gap between the races is larger and remains statistically significant even when we control for background differences.

We have presented a great deal of statistics on the gender and racial gaps in students' conceptual understanding and achievement, therefore we provide another brief summary before proceeding. In the more advanced introductory course (Physics 11a), we find that there is a statistically significant effect of pedagogy on the gender gap on the FCI. Furthermore, IE2 courses in 11a do not have a statistically significant gender gap in fitted odds of scoring high on the FCI. However, this finding does not carry over to Physics 1a and the electricity and magnetism courses, where we find that there remains a

statistically significant gender gap in performance on conceptual surveys, which disappears when we control for differences in background preparation. Additionally, we observe gender gaps on final exams and grades in almost all of the courses, which do not remain statistically significant once we control for differences in background knowledge.

In looking at the effects of pedagogy on racial differences, we do not find a statistically significant impact of pedagogy on the racial gaps in Physics 11a: IE1 courses and IE2 courses were not statistically significantly different. However, we do find differences based on type of assessment. On tests of conceptual understanding, we account for most of the outcome differences between under-represented minority students and their peers by controlling for differences in the students' background preparation. However, the racial gap persists in both the final exam scores and grades. In the next section we examine the course components in more detail, presenting the racial differences on a number of different measures.

Analysis of grades by race

To better understand the gap in student achievement by race, we present a breakdown of students' non-exam grades in Physics 1b courses, in Table 2.12. We show the average score on each component of their course grade, and the average differences between URM students and the rest of the students in the course. We tested the null hypothesis that these average differences were zero using the nonparametric Wilcoxon rank-sum test to compare the medians of the URM and nonURM score distributions. Students' scores on the first three components, reading assignments, ConcepTests, and pretests, were all based on participation or effort. Students' laboratory assignments were graded for correctness. Homework consisted of traditional quantitative problems,

assigned weekly to give students opportunities to develop more quantitative problem-solving skills. We find that most of the grades of under-represented minority students and their peers only differ, on average, by about one percentage point. On the homework assignments, however, the difference is much larger, about 3 percentage points, and are statistically significant. These results suggest that much of the racial differences in student grades can be attributed to differences in homework as well as exam scores.

Table 2.12. Breakdown of non-exam grades in Physics 1b ($N = 537$).

	URM	non-URM	difference
reading assignments	93.0	94.2	1.2 [~]
ConcepTests	86.4	87.2	0.8
pretests	86.5	87.1	0.6
laboratory assignments	96.9	98.0	1.1
homework	90.2	93.1	2.9 ^{**}

* $p < 0.05$; *** $p < 0.0001$

2.4.3 Retention in the sciences

To examine how pedagogy relates to student retention in the sciences, we tracked students who reported majoring in science at the start of the semester in an introductory physics course to determine whether they ended up graduating with a science major. Our main findings indicate that:

- The percentage of Physics 11a students who took the traditionally-taught course switched out of the science major is twice that of students in the PI-taught course;

- Furthermore, the percentages of students switching is lower in every PI course, as compared to the traditional course.

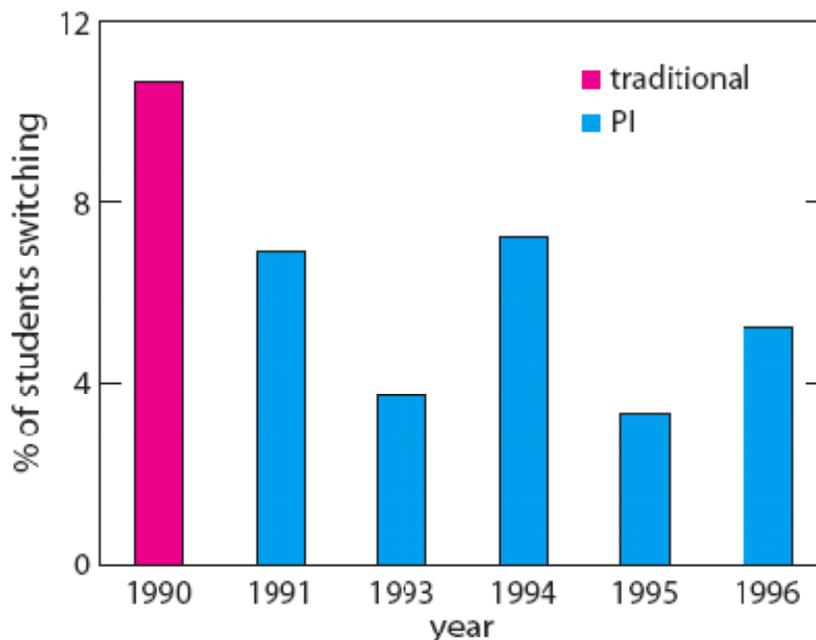
In Table 2.13, we present the sample percentages of Physics 11a students who switched from a science to a nonscience major, separated by course pedagogy. The proportion of students who were enrolled in the traditionally-taught introductory physics course and switched out of a science major is more than twice that of students enrolled in the courses taught using PI ($p = 0.02$). Furthermore, the impact of pedagogy on retention is consistent across both genders.

Table 2.13. Percentage of students who switched out of science majors, by pedagogy and by gender. ($N_{trad} = 101$; $N_{PI} = 997$).

	total	male	female
traditional	0.11	0.11	0.10
PI	0.05	0.06	0.05

As mentioned in the methods section, all courses were taught by the same instructor. The traditionally-taught course was taught in 1990, while all PI courses were taught subsequent to that course. In Figure 2.8, we show the percentages of students switching out of science majors, by year, in order to examine the fluctuations in retention across years. When comparing the traditional course with any individual PI course, the percentage of students switching out of science is more than 50% higher in the traditional course. This graph suggests that the higher percentage of students' switching out of the science major in 1990 is not simply due to yearly fluctuations.

Figure 2.8. Percentages of students switching out of science majors by year.



To better understand the role of pedagogy in student retention, we used logistic regression analyses to control for differences in student background and demographics. We account for the fact that the data are grouped by course year by using multilevel modeling. In Table 2.14, we report these fitted models, in which we predict the odds of switching out of science based on students' background and the pedagogy used in the course. We include parameter estimates, indicating associated p -values, and the goodness-of-fit statistic $-2LL$.

In Model 1, we find that whether or not a student is a freshman strongly predicts their odds of switching out of science. The fitted odds that a freshman will switch out of a science major are about 9 times the odds for upperclassmen ($p < 0.001$), on average, consistent with the findings of Seymour and Hewitt (Seymour & Hewitt, 1997). When we

control for selected student demographics in our statistical models, we find that students' fitted odds of switching out of science are not related to students' gender ($p > 0.05$). However, there is a marginally statistically significant effect of race; under-represented minority students are about twice as likely to switch out of science as their peers, on average ($p < 0.10$).

Table 2.14. Fitted logistic regression models predicting the log-odds of switching out of science majors, by selected background variables and pedagogy ($N = 1098$).

	Model 1	Model 2	Model 3	Model 4	Model 5
intercept	-3.11***	-3.32***	1.62***	0.57	1.09
freshman	2.25***	2.37***	2.35***	2.40***	2.35***
male		0.14			
URM		0.78*		0.51	
SAT math			-0.63*	-0.51	-0.58*
IE					-0.79*
-2LL	441.7	438.0	433.5	432.2	429.5

* $p < 0.05$, *** $p < 0.001$

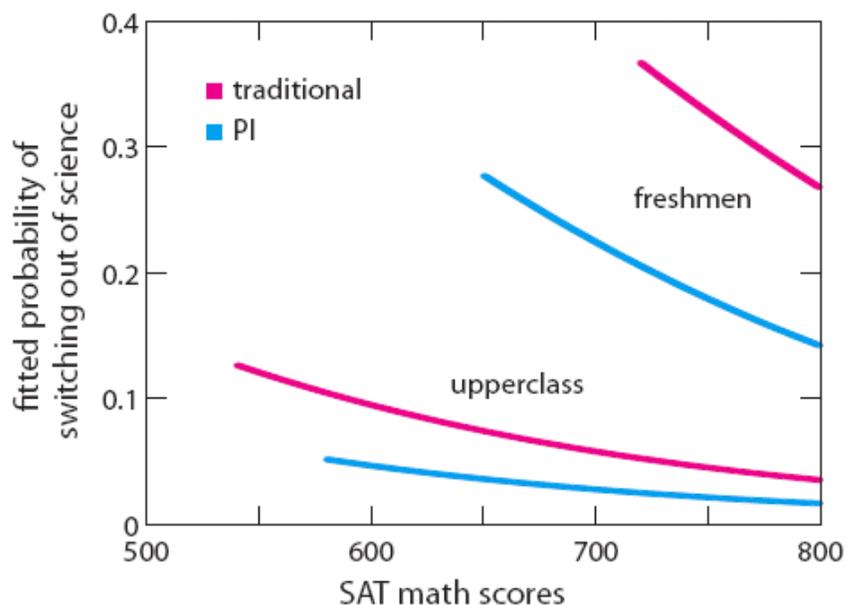
To control for mathematical background, we included students' SAT math scores as controls. Not surprisingly, we note a slightly negative relationship with SAT scores ($p < 0.05$). For every one hundred point decrease in SAT math scores, on average, the fitted odds that a student will switch out of a science major of students almost double. When we include race back in the model, we find that neither SAT math score nor race are statistically significant predictors of switching, suggesting that these two variables are potentially collinear.

Finally, in Model 5 we include our variable describing pedagogy (IE), which is a statistically significant predictor of the log-odds of a student switching out of a science

major ($p < 0.05$), even when controlling for differences in math background. On average, in the population, the fitted odds of switching out of science for students in the traditionally-taught course are more than twice the odds for students in the PI courses, consistent with the results we obtained in Table 2.13.

To graphically represent the logistic regression results, we show plots of the fitted probabilities of switching out of science in Figure 2.9. As the graph depicts, students with higher SAT math scores have a lower fitted probability of switching out of science than those with lower SAT scores, on average, in the population. Furthermore, students in a PI course tend to have lower probabilities of switching out of science than those in the traditionally-taught course, at every level of SAT math scores. Furthermore, we find that when we control for pedagogy, freshmen tend to have higher probabilities of switching out of science than upperclassmen, regardless of SAT math scores.

Figure 2.9. Fitted probability that a student will switch out of science, by pedagogy, year in school, and SAT math scores.



2.5 Discussion

2.5.1 Gender differences reduced in some IE2 courses

The Physics 11a results mirror the results published in Lorenzo, Crouch, and Mazur (2006). The effect of pedagogy on students' performance on the FCI is statistically significant; furthermore, females appear to preferentially benefit from IE2 instruction in Physics 11a. We find that there is a statistically significant gender gap in 11a IE1 course, even after controlling for selected background differences. In 11a IE2 courses, however, the difference in average FCI scores between male and female students is not statistically significant, even before controlling for differences on the FCI pretest and SAT scores. When we examine outcomes in the lower-level introductory mechanics course, Physics 1a, we find that the differences between male and female students are again statistically significant; even after controlling for FCI pretest scores, the fitted odds that males will score above 80% on the FCI posttest are twice that of female students.¹¹

The results from Physics 11a IE2 courses are promising for reducing gender differences through the use of interactive teaching. However, the results from Physics 1a suggest that more work needs to be done to address gender disparities in conceptual understanding. Although this course used IE2 methods, males still outperformed female students, on average, similar to results published at other institutions (Dockett & Heller, 2008; Kost, et al., 2009; Pollock, et al., 2007). One explanation for the differing results in our sample is that *all* students in Physics 11a were performing well; over 90% of students

¹¹ Although the Physics 11a courses used only 26 shared questions from the FCI-1 and FCI-2, the results were very similar when using the complete versions of the tests. Physics 1a analysis included all 30 questions from the FCI-2.

in the IE2 courses scored above 80% on the posttest, not leaving much room for a gender gap to occur. In contrast, only about 62% of students in Physics 1a IE2 courses scored 'high'.

In the electricity and magnetism courses, males outperform females on the end-of-semester conceptual assessment in both Physics 11b and 1b by about 0.40 standard deviations, on average. Once we control for background knowledge, however, these differences are no longer statistically significant. It is not surprising that if males are entering the course with better conceptual understanding of mechanics than females, then they would leave the course with better conceptual understanding of electricity and magnetism.

Similar to the results from University of Colorado, we do find statistically significant gender differences on final exam scores in all courses analyzed; males obtain final exam scores approximately one-third of a standard deviation higher than females, on average. However, again these differences can be explained by differences in background preparation. Similarly, males obtain slightly higher final grades, about one-quarter of a standard deviation, but these differences are no longer statistically significant once we control for background preparation.

2.5.2 Racial differences were found on exams, grades, but not on conceptual surveys

When examining the performance of Black and Hispanic students in these interactively-taught introductory physics courses, we find that their scores on the conceptual posttests are comparable to their peers with similar background preparation. However, on course exams and final grades, we find statistically significant gaps between

under-represented minority students and their peers after instruction, even when controlling for differences in background preparation. This finding is especially troubling, as course grades are important for scholarships, fellowships, admission to post-graduate schools, and provide summative feedback to students on their performance in the course.

Why do Black and Hispanic students score similarly to their peers on conceptual assessments, when controlling for background preparation, but not on course exams and grades? In this section, we consider three possible explanations for these differences, based on previous research and additional analysis of student performance: 1) stereotype threat; 2) differences in mathematics preparation; and 3) study habits.

Other studies of college grades have also found that Black students' grades are lower than their White peers with the same SAT scores, even among high-scoring students (e.g., Nettles, Thoeny, & Gosman, 1986). Furthermore, differences in income, academic background, and ability do not appear to explain racial differences in test performance and academic achievement (Aronson, 2002). Steele and Aronson (1995) hypothesized that the observed differences in achievement by race may be due to a psychological response to stereotypes about Black and Hispanic students' inferior intelligence. They found that when individuals who identify with a negatively-stereotyped group face evaluative situations in which the stereotype is relevant, such as females taking a math test or Black students taking an intelligence test, their apprehensiveness around confirming a stereotype can disrupt and depress their performance dramatically (Aronson, 2002; Steele & Aronson, 1995). This phenomenon is situational: whether or not an assessment is presented as being evaluative of ability or

preparation dramatically impacted Black students' performances on a written assessment (Aronson, 2002). In our study, the FCI and CSEM were low-stakes assessments; students only received credit for participation and therefore students may see these as less evaluative. In contrast, course exams are high-stakes and could be creating an environment in which stereotype threat would become relevant for Black and Hispanic students.

Another possible explanation for the statistically significant differences in performance on course exams, but not conceptual assessments, is that course exams require more mathematics than the conceptual surveys. When we control for FCI pretest scores and SAT math scores, we may be controlling for differences in background preparation for the FCI and CSEM, but not for the mathematics background needed for course exams. In Appendix 1, we show further results from Physics 11a using the Mechanics Baseline Test (MBT), which includes more mathematics than the FCI, but is still a multiple-choice, low-stakes, standardized assessment. We find that URM students obtain average scores that are lower than their peers, by about one-quarter of a standard deviation, when controlling for FCI pretest and SAT math and verbal scores. This disparity is statistically significant, albeit slightly smaller than the differences on final exams and grades in Physics 11a courses.

We also examine whether Treisman's (1992) work on the study habits of Black students is relevant to the racial differences we have detected in course performance. Analysis of course grades reveals that there are statistically significant racial differences on homework grades, which consisted of traditional quantitative problems. This disparity is surprising, given that the sample distribution of scores is highly skewed to the right;

most students obtain very high scores on the homework. While students are encouraged to work together in class and during section, often these discussions revolved around the concepts, not the quantitative problems on homework and exams. The cooperative problem-solving activities during the second hour of section were optional, so it's not clear whether students, particularly URM students, always received the benefit of working with their peers on quantitative problems.

With the analyses we present here, we cannot draw any conclusions about why we find racial differences on course exams and grades, but not conceptual assessments. It is likely a combination of these three factors and others not discussed here. Regardless, these results strongly show the need for additional research on the experiences of under-represented minority students in interactive classrooms. Specifically, in courses in which URM students are in the minority, with whom do URM students collaborate during Peer Instruction? Do these collaborations differ from their peers? How do different implementations of Peer Instruction affect these collaborations¹²? Furthermore, do students continue their collaborations outside of class? Answering these questions is crucial to understanding whether PI and other interactive teaching interventions are best serving under-represented minority students and determining how to modify these pedagogies to better encourage student discussion and collaboration.

2.5.3 Interactive teaching is linked to student persistence in STEM majors

The most promising finding from our analysis was the positive relationship between interactive teaching and student persistence in the sciences; all students,

¹² As found in Turpen and Finkelstein (2009), different implementations may result in different classroom norms.

regardless of race or gender, were more likely to remain in a science major through to graduation when they took an interactively-taught physics course as opposed to the traditionally-taught lecture course. As mentioned in the background review, Lichtenstein et al. (2007) found that a single positive experience in the first two years of college can encourage students to remain in an engineering major. Our finding that students who were enrolled in one interactively-taught introductory physics course were *twice* as likely as their traditionally-taught peers to persist in a science major echoes the idea that one instructor or course can have dramatic influences on students' choices.

In each of these studies, we are limited by our unique sample of students. These students were enrolled at a highly selective university, often with exceptional background preparation and family support. Regardless, our results show the need for more research and refinement of interactive pedagogies in a variety of educational settings to determine how to improve the achievement of under-represented minority students in introductory physics courses.

Chapter 3

Gender, race, and student ratings of interactive teaching

3.1 Introduction

In the last chapter we investigated the impact on achievement and retention outcomes of the use of PI and other interactive teaching techniques, specifically for under-represented groups in physics. In this section, we focus on how students respond to these teaching methods, by examining their responses to student evaluations of teaching and post-course surveys in this chapter, and their interview comments in the following chapter. Again, we are interested primarily in the responses of under-represented minority students, analyzing additional data to answer the question: how do female and racial/ethnic minority students experience these reforms in introductory physics?

In this chapter, we use quantitative methods of student responses to surveys to summarize students' evaluations of interactive pedagogies. While we want to understand particularly how under-represented students respond to these methods, we first explore how *all* students respond in their evaluations of teaching and post-course ratings of the teaching methods in a reformed introductory physics course. The literature on student evaluations of teaching (SETs) is broad; in the next section, we give only a brief review of the findings. Although a great deal of research has been conducted on validating and

understanding SETs, very little is known about how these forms function in a more student-centered, interactive classroom. Therefore, our first research question asks how students' evaluations relate to course outcomes: (1) are student learning and achievement related to their evaluations of teaching in a reformed physics course? Furthermore, the literature suggests that students' expectations about their learning—particularly their ideas about whether one obtains knowledge from an authority or builds it from their previous knowledge and experiences—can impact how they respond on SET forms. It is likely that these student expectations are particularly relevant when evaluating courses that have made use of reformed, interactive teaching methods. Our second research question addresses this issue: (2) how are student expectations about learning physics related to student evaluations of teaching in a reformed physics course?

After establishing these relationships in the courses under study, we then turn to examine how under-represented students in physics respond to these pedagogic methods, looking at both their overall course ratings and their responses to specific questions about the teaching methods used. We ask: (3) are there differences in how under-represented students evaluate the effectiveness of interactive teaching methods? In addition, we compare male and female students' responses, as well as racial/ethnic minority students and majority students' responses to the interactive pedagogies used in their introductory physics courses.

In the next section of this chapter, we give a brief review of the relevant literature on SETs—research that has been conducted primarily in traditional, didactic classrooms. We examine the previous research on the relationship between SETs and student learning and achievement. We also look closely at previous work on the multidimensional nature

of these evaluation forms and discuss what this research may mean for exploring students' evaluations of interactive classrooms. Furthermore, researchers note that the typical SET forms are designed for instructor-centered pedagogies. We use this research to inform our methods, particularly to help devise a better measure for understanding student responses to interactive teaching. The next part of our background review focuses on how student expectations and beliefs interact with their evaluation of teaching. As mentioned previously, we believe that this interaction will be especially relevant for nontraditional, student-centered classrooms, so we discuss the various measures in the literature used to assess these student expectations. Finally, we examine the research that has been conducted in reformed classrooms, which focuses on student perceptions of the teaching methods. We provide a brief review of this literature, focusing on the work that has examined gender differences (we did not find any research on the perceptions of under-represented racial/ethnic minority students).

3.2 Background

3.2.1 Are student evaluations of teaching related to student learning?

The validity of SETs—whether they are true measures of teaching effectiveness—has been the subject of much research starting as early as the 1920s (Kulik, 2001). In particular, studies have examined the relationship between SETs and subsequent learning outcomes, in an attempt to associate teaching effectiveness with student achievement. However, the results from these studies are mixed; some show a strong positive relationship between learning outcomes and student ratings, while other studies show no or a negative relationship. To synthesize these various results, Cohen (1981) provided the

first meta-analysis of research on SETs and learning, in which he reviewed the different results, determined overall trends, and examined explanations for the differences among different results. In his analysis, he looked at studies of multi-section courses with common achievement measures, comparing average scores across sections, as opposed to individual students' scores. He found a moderate to large correlation ($r = 0.43$) between the average overall course instructor rating and the average student achievement in sections. He concluded that these results not only show that "student ratings of instruction are valid index of instructional effectiveness," but that they are generalizable across contexts due to the stability of the results in different environments and with different methodologies. The differences that he did find between studies were explained by differences in instructor experience, differences in timing when the SETs were administered, and inconsistencies in evaluation across sections (e.g., if the instructor for each section graded only their students as opposed to an external grader) (P. A. Cohen, 1981).

Although Cohen determined that these results show strong support for the validity of SETs across contexts, other meta-analyses offer differing conclusions. In their review, Dowell and Neal (1982) found greater variability between studies; the relationships between SETs and learning outcomes differed widely across contexts. Furthermore, McCallum (1984) concluded that the relationship between student learning and student ratings was small; only about 6% of variability was shared in common between student performance and SETs. In a more recent meta-analysis, Clayson (2009) also found a small positive relationship, but could not reject that it was zero, in the population. Meanwhile, several other analyses confirm Cohen's original finding that SETs are related

to student learning (Feldman, 1989). These differences in meta-analytic results may be due to differences in the criteria for including studies in the analysis or differences in the calculation of individual study outcomes (Abrami, Cohen, & d'Apollonia, 1988), but also highlight the diversity of results on the validity of SETs, even when looking solely at studies comparing multiple sections within a course. As Clayson (2009) states, it can be “relatively easy to select research that reinforces a (different) point of view” about student ratings of instruction. (p. 17)

In Cohen’s original meta-analysis, he also found that the findings about the relationship between SETs and student learning did not differ by academic discipline (P. A. Cohen, 1981). In follow-up analyses, however, it was determined that many of these studies took place in introductory psychology courses and in other introductory social-science courses, making it difficult to draw conclusions about the effect of academic discipline (and other contextual factors) on this relationship (Abrami, D'Apollonia, & Cohen, 1990). “Prior analyses generally have lacked the sensitivity necessary to identify characteristics that explain a medium size effect on the relationship between ratings and achievement.” (p. 230) This lack of data from different disciplines has been echoed elsewhere (e.g., Marsh, 2007) and poses difficulty in generalizing these results to “hard” science disciplines, especially since evidence suggests that courses in these disciplines are rated lower than social science and humanities courses (Feldman, 2007). Furthermore, a more recent meta-analysis from Clayson (2009) suggests that disciplines with more “objective” measures of student achievement tend to have a smaller (if any) relationship between student ratings and student achievement.

In addition to the disciplinary threats to generalizability of SET validity results, the very nature of multi-section studies also offers reason for caution. Marsh and Roche (1997) claimed that the extant research from multi-section analyses “demonstrates that SETs reflect student learning” (p. 1189) However, multi-section courses are only a subset of courses taught at the postsecondary level and these are often introductory courses. These courses are likely more prominent at larger universities. Therefore, to say that these results are generalizable to all postsecondary courses is misleading. Adding to these objections, Weinberg, Fleisher, and Hashimoto (2007) stated that “The lack of instructor discretion in these courses leads one to question the extent to which these results will generalize to other courses” (p. 2). Furthermore, Marsh (1987) pointed out several methodological issues with multi-section studies, such as the limited statistical power from small numbers of sections. Fortunately, SET researchers advocate using multiple indicators of teaching effectiveness to assess the construct validity of SETs, including colleague and administrator observations (e.g., Marsh, 1987; Marsh & Roche, 1997), but how student ratings are related to student learning overall remains unclear.

Although not without problems, the multi-section studies are the most promising to address issues of SET validity, in which the unit of analysis must be the class or section, not the individual student. “It is the dependability of class means, rather than individual student ratings, that is of interest, and the class is the appropriate unit of analysis.” (Kane, Gillmore, & Crooks, 1976; p. 172) When deciding whether SETs are appropriate to evaluate instructors, researchers need to validate the meaning of the class averages, rather than the scores of individual students, and SET researchers generally accept the class as the appropriate unit to examine (Marsh, 2007). However, it is also

valuable to understand how individual students determine their ratings of instruction, and this insight may prove valuable in guiding future studies of SET validity (e.g., Clayson, 2007). Investigations into individual students' responses to teaching answer different research questions, emphasizing the student rather than the instructor. This research may be especially valuable to understanding the role reformed teaching methods play in student evaluations of teaching.

In summary, there are numerous studies that examine the relationship between student academic outcomes and student ratings of instruction, most of which are looking to validate the use of SETs as measures for faculty effectiveness. Although the results from various meta-analyses are mixed, there does appear to be a small, positive relationship in multi-section courses: sections in which students rate the course and instructor higher tend to have higher student achievement. However, the generalizability of these results is questionable due to the restriction of data to courses with multiple sections and when there are changes to the instructional and disciplinary context. In particular, it is unclear whether the relationship between SETs and student achievement will change when examining courses taught using reformed teaching methods, specifically in pure and applied scientific disciplines. With more emphasis directed to students' conceptual understanding and more opportunities for students to voice their ideas in class, there is reason to believe that these changes to the instructional settings may modify the relationship between SETs and student learning. While multi-section courses are needed to validate SETs in these settings (to determine whether SETs and learning gains discriminate between instructors similarly), we believe it is also valuable

to understand student ratings at the individual student level, particularly in reformed classrooms.

3.2.2 Student ratings of instruction are multidimensional

Although SETs are often characterized by a single score, often overall course or instructor rating, researchers agree that student ratings are multi-faceted (Marsh, 2007). Teaching and learning are complex activities and to be effective, an instructor must integrate several different elements, such as organizing the course, stimulating interest in the subject material, and providing feedback to students, among others, therefore it makes sense that evaluating teaching will involve several dimensions as well. Feldman (1976, 1989, 2007) defined 28 different dimensions of student ratings, categorizing items and components concerning teaching effectiveness presented in previous studies logically. Other studies have used statistical techniques, such as factor analysis, to determine different dimensions of rating empirically from student responses (Marsh, 1991). While there is theoretical agreement around the multidimensionality of student ratings, how to best evaluate instructors in light of the accepted ontology is less clear (Apodaca & Grad, 2005). Some researchers argue that global overall ratings provide the most accurate measure of teaching effectiveness (e.g., d'Apollonia & Abrami, 1997), while others contend that multidimensional profiles of SET factors best represent students' ratings of instruction (e.g., Marsh, 1987).

In light of the different dimensions suggested by SET researchers, several forms have been developed to examine student ratings of the various components of effective teaching. The Students' Evaluation of Educational Quality (SEEQ) was developed by

Marsh using items developed from literature reviews, other student evaluation forms, and interviews with students and instructors as to what comprises effective teaching (Marsh, 1982). Marsh and Dunkin (1997) claim that the content of this survey aligns with general principles and theories of adult learning. Furthermore, factor analysis has identified the nine factors on the survey consistently, providing further evidence for the multidimensionality of student ratings (e.g., Marsh & Hocevar, 1991). The Course Experience Questionnaire (CEQ) is another research-based survey developed to differentiate departments and faculties in various aspects of teaching quality (Ramsden, 1991). “The CEQ’s guiding design principle... (was) to produce quantitative data which permit ordinal ranking of units in different institutions, within comparable subject areas, in terms of perceived teaching quality.” (p. 132) CEQ items were drawn from previous surveys developed by the researchers and others, as well as student comments on open-ended items (Wilson & Lizzio, 1997). The CEQ categories were again confirmed using factor analytic techniques. Both the SEEQ and CEQ consist of 5-point Likert-scale items, asking students to agree or disagree with various statements about the course. Both surveys have undergone extensive evaluations of their reliability and validity (Coffey & Gibbs, 2001; Marsh, 1982; Ramsden, 1991; Wilson & Lizzio, 1997).

The multifaceted nature of SETs proves to be especially relevant when considering possible biases to student ratings. The most famous experiment to demonstrate bias in SETs is the Dr. Fox experiment (Ware & Williams, 1975), in which different groups of students viewed several lectures, each with different levels of content coverage and instructor charisma/expressiveness. Researchers then asked students to rate the instruction and take an examination. They found that an instructor’s expressiveness,

rather than coverage of content material, explained more of the differences in student ratings of instruction. In contrast, the coverage of content material explained more of the differences in student exam performance. However, in a re-analysis of these data to consider the multi-dimensionality of SETs, it was found that instructor expressiveness most impacted student ratings of a subset of questions related to *Instructor Enthusiasm* (Marsh & Ware, 1982). The dimension of *Instructor Knowledge* was most impacted by differences in content coverage during these presentations. While these experimental settings differ greatly from those of a college classroom, these results do suggest that looking at a dimensional profile of SETs may capture (and provide feedback on) instructors' effectiveness more accurately. These results can easily be extended to more realistic college settings and the impact of other proposed biases may be teased out of the different dimensions of SETs; for example, class size may impact ratings of group interaction, but not clarity of presentation.

While the research-based SET dimensions address some of the typical biases associated with student ratings in the literature, it is less clear how well the traditional dimensions of the questionnaires relate to courses taught using reformed teaching methods, such as Peer Instruction. Furthermore, several researchers comment on how the definitions of effective teaching underlying these student evaluations are tied to didactic, instructor-centered teaching methods (d'Apollonia & Abrami, 1997; Kolitch & Dean, 1999; McKeachie, 1997). In a PI classroom, there is less emphasis on instructor presentations, therefore the questions related to "*clear presentations*" or "*lecture organization*" may not even be relevant. In contrast, the pedagogy is more likely to impact student ratings in these courses. Therefore, more targeted items about the teaching

methods, e.g., the quality of the questions asked, the usefulness of group discussions, may be more relevant. What is considered effective teaching may be different in a PI classroom as opposed to a course using traditional lecture methods; therefore the instrument used to gather student ratings should reflect this difference.

3.2.3 Student ratings of instruction and their expectations about learning

Just as SET surveys can be biased to transmissionist ideas about learning, students' ideas about learning can impact how they rate a course. Kember and Wong (2000) present evidence of this interplay between students' ideas about learning and their perceptions of teaching quality. From 55 interviews with Hong Kong undergraduates about teaching and learning at their university, the authors developed a framework to explain why some students appreciated some forms of teaching while others did not. They suggested that there may be an interaction between students' own beliefs about learning and their perceptions of their instructors' beliefs about learning. Students' beliefs were classified along a continuum based on whether they reported taking a passive or active learning stance. Similarly, student perceptions' of the learning environment were plotted continuously from non-traditional to transmissionist. Comparing student beliefs about learning to their perceptions about different teaching environments, the authors were able to generalize student comments about what aspects of teaching they found the most and least helpful. For example, students who voiced more passive ideas about learning stated that in transmissionist teaching effective instructors are organized and systematic, while their negative comments were about lectures being rushed or "going nowhere." Meanwhile, students who reported more active stances to learning stated that

transmissionist instruction was sometimes “necessary,” but also offered negative comments for didactic teaching that blocked class discussion and participation or merely gave notes to students. Interestingly, the authors only noted negative comments from passive learners in non-traditional instruction: “The passive learners see all non-traditional teaching in a negative light because it challenges their desire to act as passive receivers of knowledge” (p. 77). The conclusions from this study are preliminary; however, they emphasize the need for further research on how student evaluations of teaching may be biased by students’ beliefs about learning.

Kember and Wong (2000) categorized students along a continuum based on their beliefs about learning; however they note that these beliefs are not fixed and immutable. Indeed, recent research on student beliefs highlight how student ideas about learning are highly context-dependent and fine-grained (e.g., Elby & Hammer, 2001; Hammer & Elby, 2002), therefore the ideas that students express during interviews may not reflect their ideas when participating in course activities or when completing teaching evaluations. Furthermore, research indicates that their ideas about learning are discipline-specific (Hofer & Pintrich, 1997; Stodolsky, Salk, & Glaessner, 1991) and therefore how students respond to different teaching methods may depend on the discipline and other contextual factors. Moreover, as McCaskey (2009) states, a student may express a preference more didactic forms of teaching, such as wanting a clear set of lectures in her science course, suggesting that she holds what Kember and Wong describe as “passive” beliefs about learning. However, in different contexts, such small-group tutorial settings, she may think this transmissionist style of teaching is less useful, even though she is in the same science course, and describe other teaching methods as being more helpful for

her learning. Therefore, to understand how student beliefs about learning interact with their response to different pedagogies and subsequent ratings of instruction, this research needs to be closely tied to the context of the specific courses and activities under study.

Researchers have used a number of different techniques to access students' expectations and beliefs about learning in a physics course, including surveys (e.g., Adams, et al., 2006; Redish, Saul, & Steinberg, 1998), interviews (e.g., diSessa, Elby, & Hammer, 2002; Hammer, 1994), and analysis of classroom behavior (e.g., Scherr & Hammer, 2009; Wutchana, Emarat, & Etkina, 2009). While surveys offer the opportunity to obtain data from a large number of students, these instruments are often very coarse and tied to the context of filling out a survey (Elby & Hammer, 2001). Interviews offer opportunities to obtain more detail about students' beliefs; however this method requires students to reflect and articulate their beliefs, which may or may not accurately represent the ideas they bring to the classroom or SET settings. Videotapes of classroom activities, particularly those that require students to talk aloud and interact with one another, offer the most promise in garnering evidence of the ideas about learning that students draw upon in the course. However, again, these student beliefs are context-specific to these activities and may not reflect the ideas they use to respond to student evaluations of teaching. Furthermore, there is evidence that students are not consistent in representing their beliefs across these different methods (Wutchana, et al., 2009). For the purposes of understanding how student beliefs interact with their ratings of teaching, it is likely that a combination of these research methods is needed. In this chapter, we rely on quantitative methods to compare students' responses to a survey on their expectations about learning in a physics course with their ratings of the nontraditional instruction used in that course.

We note that we are not necessarily capturing the beliefs that students rely on or use when participating in course activities, but rather how they respond to questions in the context of a survey about this course. In the next chapter, we take a more detailed approach to understand student ratings by examining interviews, in which students were asked specifically about their opinions on the teaching methods used in the course.

3.2.4 Student perceptions of interactive teaching

Although there is limited research on student evaluations in reformed classrooms, researchers have examined student perceptions of these interactive classroom environments. For example, several studies have been conducted on student perceptions of the use of classroom response systems, or clickers. While many of these studies emphasize the technology rather than the pedagogy, results from a variety of disciplines have shown positive student responses to the use of this technology in the classroom (e.g., Crossgrove & Curran, 2008; Hoekstra, 2008; Rice & Bunz, 2006). One study that focused on the pedagogy was conducted in a first-year engineering mechanics course. In this study, researchers examined students' responses to two different forms of interactive teaching: PI and class-wide discussions as promoted by Dufresne, et al. (1996). In both surveys and focus group interviews, students reported that they found the procedure of PI, including initial individual vote and subsequent peer discussion, helpful to their understanding of the material. Moreover, students found PI more beneficial to their learning than class-wide discussions. Although these results are encouraging for the use of PI, other studies have shown that student perceptions may depend on academic

discipline, implementation, and institutional context (S. W. Draper & Brown, 2004; Turpen & Finkelstein, submitted).

Many of the papers reporting on student perceptions of interactive teaching have mainly focused on the averages of *all* students' responses; only few have reported on the impact of gender in student response to these methods. Studies on clicker use which have reported on gender have primarily found no differences between males' and females' ratings on the use of this technology (Edens, 2008; Freeman & Blayney, 2005; Rice & Bunz, 2006). However, these studies focus on the technology, rather than the pedagogy—which in some cases did not involve peer discussions or even conceptual questions. Studies from PER on reformed teaching pedagogies, which are more aligned with the principles of PI even though they do not use the same tools (clickers), suggest that there may be gender differences in response to the interactive, collaborative nature of these pedagogies (Heller & Hollabaugh, 1992; Laws, et al., 1999). Therefore, to gain a better understanding of how students respond to PI, we should look at gender and other student demographics to see if there are differences in how students evaluate this method.

In summary, while there has been a great deal of research conducted on student evaluations of teaching in traditional classroom settings, how students evaluate reformed, interactive, constructivist teaching is less well-known. Furthermore, the traditional forms used in teaching evaluations have been shown to be biased to lecture-based, instructor-centered didactic forms of teaching. To address these issues, we provide an in-depth investigation of SETs in courses taught using PI and other interactive teaching methods. We use both the conventional forms used in student evaluations of teaching as well as develop our own 'dimension' of questions examining the teaching effectiveness of

interactive teaching methods. Using these tools, we investigate the relationship between SETs and student learning/achievement outcomes, as well as their student expectations about learning physics. Finally, we examine how under-represented groups respond to and rate these teaching methods to better understand how these students experience interactive classrooms in introductory physics. For clarity, we restate the research questions we explicated in the introduction section for easy reference for the reader:

- (1) Are student learning and achievement outcomes related to their evaluations of teaching in a reformed physics course?
- (2) How are student expectations about learning related to student evaluations of teaching in a reformed physics course?
- (3) Are there differences in how under-represented students evaluate interactive teaching methods?

3.3 Methods

3.3.1 Course information

We conducted this study in two different courses at Harvard University over three years: Physics 1b in spring 2005 and 2006 and Physics 11b in spring 2007. Both Physics 1b and 11b were the second-semester courses in a one-year sequence in introductory physics. Physics 1b was designed as a general introduction to physics for students majoring in the life sciences, satisfying pre-medical school requirements and requiring only limited use of calculus in the course. Physics 11b is an intermediate-level introductory physics course, also designed for non-physics majors, but involving more

calculus and concurrent enrollment in multivariable calculus. Both courses covered topics in electricity and magnetism, circuits, and optics using a draft research-based textbook written by the course instructor.

The courses were taught using a variety of interactive teaching methods by the same instructor with extensive experience using these teaching methods. Before class, students were required to read the textbook and answer two short, conceptual questions about the material. Students were also asked what part of the reading they found most difficult or confusing. Using this feedback about students' level of understanding, the instructor tailored the lectures to specifically address the concepts that students found most challenging. During the large lecture meetings, the instructor used PI (Mazur, 1997), first reviewing the material by giving short presentations and then asking several ConcepTests during the 1.5 hour meetings. Students received participation credit for responding to these questions in-class, but did not receive extra credit for answering correctly. The course also included two-hour optional sections run by graduate teaching assistants each week. During the first half of the section, students worked in small groups on *Tutorials in Introductory Physics* (McDermott, et al., 2002). During the second hour, the teaching assistants led cooperative problem-solving sessions, to give students practice with quantitative problems and help them with their homework. A more detailed description of the course methods can be found in chapter 2.

3.3.2 Measures

In this section, we discuss the following four categories of measurement tools we used to address our research questions:

- Learning and achievement measures
- Student evaluations of teaching
- Survey on student expectations about learning physics
- Survey on student rating of teaching methods

Learning and achievement measures

In all three courses, we collected data related to both students' conceptual understanding of the topics covered in the course and their achievement (grades) in the course. We did not administer a pretest of student understanding of electricity and magnetism; we did not believe that test scores would have much meaning as the terms and concepts are often too unfamiliar to students who have not taken a relevant physics course. Instead, we administered the Force Concept Inventory (FCI; Hestenes, et al., 1992) during the first two weeks of class. The FCI assesses students understanding of Newtonian mechanics, which is a building block for many concepts in electricity and magnetism. At the end of the semester, two weeks before the final exam, we measured students' conceptual understanding of electricity and magnetism using the Conceptual Survey of Electricity and Magnetism (CSEM; Maloney, et al., 2001). Students completed both the FCI and CSEM online. Students had one hour to complete the assessments, without the use of the textbook or other materials. The instructor awarded credit for participation in completing these assessments, but students did not receive extra credit for correct answers.

We also collected data on student achievement in the course, focusing on students' performance on the final exam and their final grade. When grouping Physics 1b students across semesters, we transformed student scores into z-scores, using the mean and standard deviation from each semester. In both Physics 1b and 11b, the exams included both quantitative and conceptual problems.

Student evaluations of teaching

With help from the Office of the Registrar at Harvard University, we matched student demographics and achievement in Physics 1b and 11b with individual students' official end-of-semester evaluations, known at Harvard University as the "Q evaluations." We focus on students' response to the item asking students to "evaluate this course overall," which we refer to as their *overall evaluation*.

In all three semesters, students completed the Harvard Q evaluations online. Students had several weeks to complete this optional questionnaire and the evaluation tool was closed just before the scheduled final exam. The results from the evaluations were published every year in the printed Harvard Q Guide and online. The evaluation forms in spring 2005 and spring 2006 were very similar, while changes were made in 2007.

Survey of student expectations about learning physics

To answer our second research question, we surveyed Physics 11b students about their expectations about learning physics in spring 2007. During the first two weeks of class, students were asked to complete a pre-course survey that included 48 questions about their expectations for the course, their expectations about learning physics, their interest in the subject and background in physics. Most of these questions were taken

from the second version of the Maryland Physics Expectations Survey (McCaskey, 2009). We also asked students about their motivation for taking the course and how they approach learning, with several items based on the Lancaster Inventory of Approaches to Learning (Entwistle, Hanley, & Hounsell, 1979); all survey questions are presented in Appendix 2. Most of the items asked students to rate their level of agreement to a statement on a five-point Likert scale. Other questions were more elaborate; for example, several questions described a conversation, and then asked students with which student they agreed, on a five-point scale. We transformed all of the answers such that a score of ‘five’ represented an expert-like response and a score of ‘one’ a more novice-like response. Students completed the survey online and received a small amount of credit for participation.

Instead of considering students’ responses to each question individually, we used hierarchical cluster analysis to group together similar questions from the pre-course MPEX and approaches to learning survey questions. Cluster analysis is a statistical technique that groups items (survey questions) together based on how similarly students responded to them. There are a number of ways to perform cluster analysis; we used the VARCLUS procedure in SAS. We then examine these clusters of items, as well as the content of the items in each, to best categorize the survey questions in our final groupings.

While many surveys on student epistemologies employ factor analysis to determine the coherence of different dimensions they presume the survey items to measure, the developers of MPEX-2 take a principled stand against these methods, claiming that they ignore the context-sensitivity of student expectations and presume that

students will answer all items for a given dimension similarly (McCaskey, 2009). In our analysis, we chose to use cluster analysis not to confirm predetermined dimensions, but to obtain clusters of items that students in our sample answered similarly. We are not claiming to measure a given dimension of students' beliefs with these clusters or to make claims about the underlying structure of student beliefs impacting responses, but we simply use this as a way to identify interesting groups of survey questions to which students responded similarly. We then use a linear combination of students' responses to the items in each cluster to represent their net response to the cluster in subsequent analyses.

From inspection of the cluster-analytic results, we grouped the survey items into six clusters, with one singleton question that was relatively independent of the other items. Student responses to items within each cluster were then examined again separately using principal components analysis (PCA). In PCA analyses, the first principal component is the linear composite of responses to the items with that cluster that contains the most of the original item-level variability in responses. To determine how each item contributes to the composite, we can inspect its loading on the first principal component. To obtain a composite score on all the items in a given cluster, we can use the score on the first principal component to provide a 'weighted' average of student responses to those items (which also allows items with reverse scales to contribute negatively to the composite). All-in-all, these statistical procedures provide a way to group items that students responded to similarly and provide us with an optimized linear combination of their responses to these items.

In Table 3.1, we present our categorization of the pre-course survey questions, guided by these statistical techniques. We describe each cluster briefly, noting the common themes that stand out in each group of questions. The first cluster consists of eight questions, most of which are related to student expectations about what they will need to know to be successful in this introductory physics course. Some of the items refer to the role of formula-matching in physics, while others refer to memorization. A theme underlying all of these questions is how actively engaged with the material the students expect (or prefer) to be. Students who score ‘high’ on this cluster may expect and appreciate the need for deep understanding of concepts and recognize that this understanding comes from active engagement of the student with concepts, as opposed to formula matching, memorization, and listening to lecture. This cluster of questions, which we refer to as *conceptual understanding* appears to include both “independence” and “concepts” dimensions from the MPEX (McCaskey, 2009; Redish, et al., 1998).

Cluster 2 contains only two questions, both directly related to student expectations of the classroom meetings with the professor. Students who agree with these statements may expect their physics professor to use classroom time to cover all of the content and show how to do problems that they need to know for the exam. This cluster appears to capture the “authority” dimension of the MPEX; we call this variable *role of professor*.

Cluster 3 contains seven questions, four of which specifically involve the importance of formulas and specific methods in physics problem-solving. Two other questions, “*Knowledge in physics consists of many pieces of information, each of which applies primarily to a specific situation*” and “*A good physics textbook should show how the material in one chapter relates to the material in other chapters. It shouldn’t treat*

Table 3.1. Clusters of items from the pre-course survey administered in Physics 11b.

Cluster 1: conceptual understanding

“Problem solving” in physics basically means matching problems with facts or equations and then substituting values to get a number.

Two students are talking about their experiences in class.

Meena: Our group is really good, I think. We often spend a lot of time confused and sometimes never feel like we have the right answer, but we all listen to each other's ideas and try to figure things out that way.

Salehah: In our group there is one person who always knows the right answer and so we pretty much follow her lead all the time. This is great because we always get the tasks done on time and sometimes early.

With whom do you agree?

Let's say a student has limited time to study, and therefore must choose between the following options. Assuming the exam will be a fair test of understanding, and assuming time pressure during the exam isn't an issue, which option should the student choose?

a) Learning only a few basic formulas, but going into depth with them

e) Learning all the formulas from the relevant chapters, but not going into as much depth.

Roy and Theo are working on a homework problem.

Roy: 'I remember in the book it said that anything moving in a circle has to have a centripetal acceleration.'

Theo: 'But if the particle's velocity is constant, how can it be accelerating? That doesn't make sense.' Roy: 'Look, right here, under 'Uniform Circular Motion' - here's the equation $a = (v^2)/r$. That's what we need for this problem.'

Theo: 'But I know that to have an acceleration, we need a change in velocity. I don't see how the velocity's changing. That equation doesn't seem right to me.'

If you could only work with one of them, Roy or Theo, who do you think would be more helpful?

I learn best in classes during which I participate and ask lots of questions.

I prefer classes during which the professor lectures and recites all of the material to be tested on the exam.

The main point of seeing where a formula comes from is to learn that the formula is valid and that it is OK to use it in problems.

Some people have ‘photographic memory’, the ability to recall essentially everything they read. To what extent would photographic memory give you an advantage when learning physics?

Table 3.1 (Continued)

Cluster 2: role of professor

If physics professors gave really clear lectures with plenty of real-life examples and sample problems, then most good students could learn those subjects without having to spend a lot of time thinking outside of class.

I consider classroom meetings to be the place where I will get all the information I need for the exam.

Cluster 3: physics problems

The most crucial thing in solving a physics problem is finding the right equation to use.

To really help us learn physics, professors in lecture should show us how to solve lots of problems, instead of spending so much time on concepts, proofs of general equations, and one or two problems.

Group work in physics is beneficial only if at least one person in the group already understands and knows what they are talking about.

When solving problems, the key thing is knowing the methods for addressing each particular type of question. Understanding the “big ideas” might be helpful for specially-written essay questions, but not for regular physics problems.

To understand physics, the formulas (equations) are really the main thing; the other material is mostly to help you decide which equations to use in which situations.

In this course, adept use of formulas is the main thing needed to solve physics problems effectively.

Knowledge in physics consists of many pieces of information, each of which applies primarily to a specific situation.

Tracy: A good physics textbook should show how the material in one chapter relates to the material in other chapters. It shouldn't treat each topic as a separate 'unit', because they're not really separate.

Carissa: But most of the time, each chapter is about a different topic, and those different topics don't always have much to do with each other.

The textbook should keep everything separate, instead of blending it all together.

With whom do you agree?

Table 3.1 (Continued)

Cluster 4: sense-making

In this course, I do not expect to understand equations in an intuitive sense; they just have to be taken as givens.

When doing practice problems for a test or working on homework, if I came up with two different approaches to a problem and they gave different answers, I would not worry about it; after finding out the right answer, I'd just be sure to avoid the incorrect approach.

Often, a physics principle or theory just doesn't make sense. In those cases, you have to accept it and move on, because not everything in physics is supposed to make sense.

Tamara just read something in her physics textbook that seems to disagree with her own experiences. But to learn physics well, Tamara shouldn't think about her own experiences; she should just focus on what the book says.

Although physical laws may apply to certain simple situations like we see in class and lab, they have little relation to what I experience in the real world.

I like to come to class and just write down notes without thinking much about what I'm writing. I'll go over my notes when I do homework or study for the exam.

Cluster 5: self-confidence

If I don't remember a particular equation needed for a problem in an exam I can probably figure out an (ethical!) way to come up with it, given enough time.

If a problem on an exam does not look like one I've already done, I don't think I would have much of a chance of being able to work it out.

When handing in a physics test, you can generally have a correct sense of how well you did even before talking about it with other students.

Table 3.1 (Continued)

Cluster 6: real world

When learning a new physics topic it's important to think about my personal experiences or ideas and relate them to the topic being analyzed.

Physics is related to the real world, but I can understand physics without thinking about that connection.

My grade in this course will be primarily determined by how familiar I am with the material. Insight or creativity will have little to do with it.

Consider the following question from a popular textbook:

'A horse is urged to pull a wagon. The horse refuses to try, citing Newton's 3rd law as a defense: The pull of the horse on the wagon is equal and opposite to the pull of the wagon on the horse. "If I can never exert a greater force on the wagon than it exerts on me, how can I ever start the wagon moving?" asks the horse. How would you reply?'

When studying for a test, what best characterizes your attitude towards studying and answering questions such as this?

- a) Studying these kinds of questions isn't helpful, because they won't be on the test.
 - b-d)
 - e) Studying these kinds of questions is extremely helpful, worth a whole lot of my study time.
-

During class meetings, I think about the material as I am taking notes, and think of questions to ask the teaching staff either during or after class.

Learning physics will help me understand situations in my everyday life.

It wouldn't matter if I didn't get my homework returned to me as long as I knew which questions I got wrong and I had the solutions to study.

Singleton

Several students are talking about group work.

Carmela: 'I feel like explaining something to other people in my group really helps me understand it better.'

Juanita: 'I don't think explaining helps you understand better. It's just that when you can explain something to someone else, then you know you already understood it.'

With whom do you agree?

each topic as a separate ‘unit,’ because they’re not really separate,” do not appear to be tied directly to the other questions, but address students’ ideas about the coherence of physics knowledge. However, if a student views physics problem-solving as formula matching, then they *would* be considering physics knowledge as discrete “pieces of information,” specifically as formulas. Regardless, these two questions about coherence have the lowest correlation with the other questions and do not contribute as much to the composite generated from PCA. One other item, “*Group work in physics is beneficial only if at least one person in the group already understands and knows what they are talking about,*” also appears to be unrelated to the other questions. However, interviews with students at the start of the semester revealed that most students only worked together when doing problem sets. The question could be interpreted such that students who agreed with this statement thought working on problem sets together is beneficial if only one person knows the correct methods. This cluster, which we will call *physics problems*, appears to mix the dimensions of “concepts” and “coherence” from the MPEX, specifically as they relate to physics problem-solving.

Items in cluster 4 all refer to the role of sense-making in learning physics. Students score ‘high’ on this cluster when they believe that they need to make sense of physics with their own experiences and intuition to learn the material. There is also an aspect of reconciliation in these questions – students that score ‘high’ recognize the need to reconcile two different approaches to a problem or how a physics principle may go against their ideas about the world. This cluster is a mixture of the ‘independence,’ ‘concepts,’ and ‘real world’ clusters from the MPEX. We refer to this variable as *sense-making*.

The items in cluster 5 all involve expressions of student confidence in their abilities on physics exams; we therefore refer to this variable as *self-confidence*. The first two questions ask about their opinions on their ability to perform well on a physics exam. The last question in the cluster, “*When handing in a physics test you can generally have a correct sense of how well you did even before talking about it with other students*”, asks students how well they believe they can monitor their own performance.

Cluster 6 contains seven questions that are similar to those in clusters 1 and 4 (*conceptual understanding* and *sense-making*). However, four of these items refer more directly to how physics relates to the real world. The other questions also deal more explicitly with role of the student in understanding the material, whether to do well the student needs to make connections with their previous knowledge and experiences (presumably from the ‘real world’) or just be ‘familiar’ with the material the professor presents. This last variable is not surprisingly called *real world*.

By compositing responses to the questions in these six clusters, we obtain six variables that measure different aspects of student expectations about learning physics, for each student: *conceptual understanding*, *role of professor*, *physics problems*, *sense-making*, *self-confidence*, and *real world*. We then examine the relationship between scores on each of these composite measures and students’ rating of the teaching methods, which we describe in the following section.

Survey of students’ ratings of teaching methods

Our last measurement instrument allowed for a more detailed analysis of student evaluations of the teaching methods. Using previous surveys as a guide, we developed and administered an additional survey that we administered to students in Physics 11b.

During the final two weeks of the course, we asked students 42 questions about their opinions of different features of the course, including the assessment, clarity of course goals, and teaching methods used. Most survey items were taken from Course Experience Questionnaire (CEQ; Wilson & Lizzio, 1997) and the Students' Evaluation of Educational Quality (SEEQ; Marsh, 1982); we also developed five additional questions that asked students about the specific methods used in Physics 11b. Several survey questions from each category were pilot-tested in student interviews. All questions are listed in Appendix 2. We sought student responses to each question on a five-point Likert-scale. In our analysis, we reversed the scale on some questions so that a score of 'five' always represents a favorable response to the methods and a score of 'one' represents an unfavorable response. When presenting analyses of responses to these items, we will denote items with reversed scales using the letter 'R'.

To develop a composite measure for student opinions about the teaching methods used in the course, we carried out exploratory hierarchical cluster analysis on the responses to the post-course survey items, excluding the items that asked for students' rating of the course overall and difficulty. In Table 3.2, we show the items that comprised the cluster of interest for this project. This cluster includes: all of the questions designed to assess student opinions on the methods used, two questions that asked about the academic value of the course, one question about the quality of teaching and another question assessing student engagement. We anticipated that student responses to the teaching methods were linked to their responses to questions about the academic value of the course. The item about student engagement is also related: "*This course stimulated me to ask more questions.*" The statement from the CEQ on quality of teaching ("*Our*

instructor is extremely good at explaining things to us.”) is a more surprising addition to this cluster, as the process of Peer Instruction focuses on having students explain to other students, not just the transfer of information from instructor to student. It may be that students who responded more favorably to PI did not feel that the decreased emphasis on instructor explanations was inadequate, while those who did respond less favorably to PI noticed the difference from lecture-style classes and responded less favorably to this statement.

Table 3.2. Items included in *methods* cluster. “R” indicates that the scale on that item was reversed.

Methods cluster items
I found this class intellectually challenging and stimulating.
Answering ConcepTests in class helped me learn the material.
This course stimulated me to ask more questions.
Our instructor is extremely good at explaining things to us.
Discussing the answers to in-class questions with my peers was valuable for my learning.
I felt the methods used in my first semester course, Physics 11a/Physics 15a, were more helpful than those used in this course to help me learn physics in an in-depth manner. R
Reading the book before class did not help my understanding of the material. R
I feel that traditional lectures, with clear presentations of the material I need to know for the exam, are more beneficial for my learning than the method used in 11b. R
In Physics 11b, I learned something which I consider valuable.
I would want to see other science courses taught in a similar way to how this course was taught.

Using responses to the items in this “methods” cluster, we again performed PCA to confirm our grouping. We found similar loadings for all items on the first principal component, suggesting that each item contributed similarly to the composite. For simplicity, and to retain the original five-point scale, we then averaged student responses to these items to form our composite dependent variable, *methods*.

3.3.3 Description of the Sample

In Table 3.3, we present descriptive statistics on students in Physics 1b (spring 2005 and 2006) and Physics 11b (spring 2007), including the total number of students, the percentage of male students, and the racial breakdown of students enrolled in the course. Note that in Physics 1b, the majority of students were female (68%); in Physics 11b there were approximately equal percentages of males and females. Most of the students in these courses were Asian-American or White. We also display information about students' declared majors, which we obtained from the Office of the Registrar. In Physics 1b, just over half of the students were non-physics science majors; the other students were non-science majors. Almost all (92%) of Physics 11b students were science majors. Most of the students in Physics 1b were juniors (67%), while Physics 11b was more evenly split between sophomores (38%) and juniors (46%).

Table 3.3. Univariate descriptive statistics on student demographics and majors for all students enrolled in Physics 1b (spring 2005 and 2006) and Physics 11b (spring 2007).

	2005	2006	2007
<i>N</i>	164	169	130
male	35%	30%	54%
Asian-American	27%	17%	26%
Black	9%	11%	6%
Hispanic	8%	7%	7%
Foreign	10%	7%	15%
Native American	1%	2%	2%
White	38%	49%	31%
Other/unknown	7%	7%	13%
Physics major	0%	1%	1%
Other science major	52%	57%	92%
Nonscience major	48%	42%	7%

Not all students chose to complete the FCI and CSEM. In Table 3.4, we show the response rates on these two measures of student learning, as well as the response rate for the Harvard Q evaluations. SETs are notorious for their low response rate, especially when administered online (e.g., Dommeyer, Baum, Hanna, & Chapman, 2004; Layne, Decristoforo, & McGinty, 1999). Therefore, we only have complete matched datasets for approximately half of the students in the course.

Table 3.4. Sample response rates on measures of student learning and student responses to the university-administered end-of-semester student evaluations for Physics 1b (spring 2005 and 2006) and Physics 11b (spring 2007).

	<i>N</i>	FCI	CSEM	SET	all data
2005	164	95%	70%	69%	52%
2006	169	99%	91%	54%	46%
2007	130	78%	93%	62%	51%

To answer our research questions in more detail, we focus on responses to the Physics 11b pre-course survey on student expectations about learning and post-course survey about the teaching methods. The response rates for these surveys are 76% and 72%, respectively. Out of the 130 students in the course, only 73 students completed both surveys. In Table 3.5, we present descriptive statistics on the FCI scores, CSEM scores, final exam scores, and final grades for students who completed both surveys (no miss) as well as for students missing data on one or both of the surveys (missing survey). We present the number of students, mean scores, and their standard deviations, in each group and we compare their average differences as effect sizes¹³ (ES). To test whether the two

¹³ Effect size is estimated by subtracting the female mean from the male mean and dividing by the pooled standard deviation.

groups were drawn from the same population, we performed one of two statistical tests, depending on whether the distribution of scores resembled a normal distribution. For final exam, we performed a paired t-test, comparing the mean scores in the two groups. In parallel analyses, we performed the Kruskal-Wallis rank sum test, to compare the median scores on the FCI, CSEM, and final grades. The p -values resulting from these tests are shown in the last column. Note that all the achievement measures at the end of the course reveal statistically significant differences between students who completed the survey and those who did not. These differences indicate that our data are not missing completely at random.

Table 3.5. Comparison of FCI scores, CSEM scores, final exam scores, and final grade between students who completed both the pre-course survey on student expectations about learning physics and the post-course survey on student ratings of interactive teaching methods.

	took both surveys			missing one+ survey			difference	
	N	mean	SD	N	mean	SD	ES	p
FCI	69	24.7	4.9	33	23.8	5.1	0.18	0.47
CSEM	71	25.7	3.7	50	23.2	5.3	0.54	0.01
final exam	73	79.6	10.9	57	75.5	12.8	0.34	0.05
final grade	73	91.2	5.4	57	87.8	7.7	0.51	0.006

note: p -values were obtained using a two-tailed t-test for the final exam score outcome and the Kruskal-Wallis rank-sum test for the FCI, CSEM scores and final grades.

3.3.4 Analytic methods

We use the student evaluation data from the Harvard Q evaluation to answer our first research question. We estimate the correlation coefficients between students' *overall evaluation* of the course and students' scores on the FCI, CSEM, final exam, and final

grades. Because the scale of the overall evaluation measure is coarse (i.e., five-point scale), the correlation coefficients that we obtain are attenuated (e.g., Martin, 1978). Therefore, we apply the correction factor proposed by Peters and van Voorhis (1940) to account for the downward bias (Aguinis, Pierce, & Culpepper, 2009).

To answer our second research question, we focus on the composites generated by cluster analysis of the surveys administered in Physics 11b, and examine how student expectations about learning physics relate to student evaluations of interactive teaching. In each case, we estimate bivariate correlations between each composite from the pre-course expectations and the score on the *methods* composite. We use those expectation variables that have statistically significant relationships with student ratings of the methods in multiple regression analyses to address the third research question.

To determine whether males and female rate these courses differently, we first conducted two-tailed t-tests, comparing male and female students' *overall* evaluations of the course from the Harvard Q evaluation. Although the Q evaluations were administered all three semesters, we only administered the surveys on student expectations about learning and their evaluations of interactive teaching methods for students in Physics 11b (spring 2007). We perform additional analyses with this data, using multiple regression analyses to examine whether the relationship between student expectations about learning and student evaluations of the teaching methods differ for males and females. Controlling for student expectations, we first examine whether there are statistically significant differences between males and females in rating of the course methods. We then include a two-way interaction variable to determine whether the relationship between student expectations and student ratings differed by gender.

The analyses described in the previous section suggest that ignoring students without complete datasets could produce misleading conclusions; the “complete case” sample is not representative of the entire course. Furthermore, this “listwise-deletion” method of dealing with missing data has been widely discredited in the literature (e.g., Little & Rubin, 1987). In addition to list-wise deletion, other ad-hoc methods have been proposed to deal with missing data, each with severe statistical drawbacks (e.g., Schafer & Olsen, 1998). Mean substitution, which replaces the missing values with the average value across the sample on that variable, produces biased results, and artificially reduces the variance for the variable and diminishing its relationships to other variables. Furthermore, replacing the missing values using regression substitution, or any other method that imputes the values only once does not account for the error introduced by estimating the missing data. Therefore in our statistical analyses we use the method of multiple imputation (Rubin, 1987), to replace the missing values at random, multiple times, and then to account for the uncertainty by pooling the multiple slope estimates obtained in our regression analyses to examine the relationships between gender, race, student expectations about learning, and their ratings of the teaching methods.

We use the method of multiple imputation to create several complete datasets using appropriate randomly imputed values for the missing data in each variable, based on the non-missing data we possess for individual students, so even if the data are not missing completely at random (i.e., the probability of missing data depends on one of the other variables such as final grade) we account for this bias (Wayman, 2003). Once we have these multiple datasets, we conduct identical regression analyses in each and pool the estimates. To obtain the overall estimate of a regression parameter, for instance, we

average the individual estimates from each of the m imputed datasets. The variance (T) of the estimated regression parameter is estimated by the formula given by Rubin (1987):

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B$$

where \bar{U} is the average variance within each dataset and B is the variance between datasets. Because the total variance accounts for the differences between imputed datasets, the inferences we draw from these analyses—such as standard errors and p -values—account for the uncertainty from imputing the missing data. This method of analysis has been shown to work remarkably well with relatively few imputed datasets (Rubin, 1987) and has proven to be a “principled and statistically defensible” approach to dealing with missing data (Schafer & Olsen, 1998).

3.4 Results

3.4.1 Relationship between student ratings of interactive teaching and learning outcomes

In this section, we address our first research question, exploring the relationship between student evaluations and learning. Because our results are mixed, based on course, we provide a brief summary of our findings here to help guide the reader through our results:

- While the scores on the conceptual surveys differed by course, students’ average *overall rating* remained consistent at about 2.5 on a scale of 1–5 (5 being highest).

- In the lower-level introductory course (Physics 1b), we find that the estimated correlations between student ratings and course achievement differ by semester. The only correlation that remains consistent from semester to semester is between students' CSEM scores and their *overall rating* from the SET.
- In the more advanced introductory course (Physics 11b), we find that achievement variable that has the largest correlation with student ratings is student performance on the FCI, administered at the beginning of the semester.

In Table 3.6, we present summary statistics for students' performance on the conceptual tests and course achievement as well as students' overall rating of the course on the university-administered end-of-semester evaluations (SETs). We display the number of students, mean, and standard deviation on each of these measures in the lower-level introductory course, Physics 1b (spring 2005 and 2006), and in the more advanced introductory course, Physics 11b (spring 2007). We only include student data for those students who completed the end-of-semester evaluations.

Students in Physics 1b obtained average pre-course FCI scores that are similar to previously-reported scores obtained by Harvard University students after a semester of traditional mechanics instruction in Physics 1a (Crouch & Mazur, 2001). Not surprisingly, students in Physics 11b, a more advanced introductory course, enter the course with higher FCI scores. These scores also correspond to previously-reported average posttest scores in traditional and moderately interactive courses in Physics 11a

(Crouch & Mazur, 2001). Physics 1b students' post-course CSEM scores are comparable to those obtained by physics majors and graduate students, while the Physics 11b average score is comparable to the average score obtained by two-year college professors (Maloney, et al., 2001). The final exam was different in each semester; therefore the average scores are not comparable even within Physics 1b. We do find similar average responses to students' overall rating. In each semester, students rated the course overall at about 2.5 on a scale from 1–5.

To examine the relationship between students' evaluations of the course and the learning and achievement outcomes, we display the corrected estimates of the correlation coefficients between student ratings and student achievement in Table 3.7. In the spring of 2005, Physics 1b students' overall ratings of the course were slightly correlated with students' CSEM scores ($r = 0.20$); this correlation was marginally statistically significant ($p = 0.08$). On all other measures, the estimated correlation coefficients were very low and not statistically significant ($r = -0.01$ – 0.05). We find stronger correlations between course achievement and students' overall ratings in the spring 2006 semester. Students' CSEM scores are again only slightly but statistically significantly correlated with rating ($r = 0.29$, $p = 0.02$). However, unlike spring 2005, we find that final exam scores and final grades are also statistically significantly correlated with students' overall ratings ($r = 0.24$ and $r = 0.31$, respectively).

Table 3.6. Summary statistics on students' *overall rating* for the end-of-semester teaching evaluations and students' performance on conceptual tests and course achievement.

	2005			2006			2007		
	<i>N</i>	mean	SD	<i>N</i>	mean	SD	<i>N</i>	mean	SD
FCI	101	21.53	5.17	83	21.77	5.72	70	25.13	4.50
CSEM	88	22.19	4.86	77	22.23	4.86	78	25.69	3.73
final exam	104	63.89	12.65	83	71.16	11.65	82	82.13	9.63
final grade	104	85.76	5.90	83	88.45	5.23	82	91.95	4.97
overall rating	104	2.49	1.12	83	2.43	1.27	82	2.55	1.26

120 Table 3.7. Estimated bivariate correlation coefficients relating students' *overall rating* for the end of semester teaching evaluations and students' performance on conceptual tests and course achievement for Physics 1b (spring 2005 and 2006) and Physics 11b (spring 2007).

	2005		2006		2007	
	<i>N</i>	<i>r</i>	<i>N</i>	<i>r</i>	<i>N</i>	<i>r</i>
FCI	101	0.02	83	0.12	70	0.39**
CSEM	88	0.20~	77	0.29*	78	0.22~
final exam	104	-0.01	83	0.24*	82	0.19
final grade	104	0.05	83	0.31**	82	0.19

~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

note: these estimates were corrected for coarseness of scale on *overall rating*, by dividing each estimate by the recommended value for 5-point scales (0.943) [Peters and van Voorhis, 1940].

We transformed students' final exam scores and final grades to combine both semesters of Physics 1b. Consistent with our semester-by-semester results, we find a statistically significant correlation between CSEM scores and overall rating ($r = 0.24, p = 0.003$). We also find a statistically significant relationship between final grades and overall rating, however the magnitude of the estimated correlation coefficient is lower ($r = 0.17, p = 0.03$).

Although we did not find a statistically significant relationship between students' overall ratings and their pre-course FCI scores in Physics 1b, this is the only statistically significant correlation we obtained in analyses of the Physics 11b data ($r = 0.39; p = 0.002$). We find a marginally statistically significant relationship between CSEM scores and students' overall rating, but not with the other achievement variables.

For a more specific measure of students' ratings of the teaching methods used in the course, we administered a post-course survey to students in Physics 11b and generated a composite variable by averaging student responses to ten questions. In Figure 3.1, we show the histogram of students' ratings of the interactive teaching methods used in the course. We find that these responses are strongly correlated with students' overall ratings obtained on the university-administered student evaluations ($r = 0.89, p < 0.0001$).

In Table 3.8, we show estimated correlation coefficients that summarize the relationship between course achievement and this new variable measuring students' ratings of the teaching methods used in Physics 11b. We again find that students' performances on the conceptual surveys are slightly, yet significantly, correlated with

student ratings. We do not find statistically significant relationships between course grades and student ratings.

Figure 3.1. Histogram of student ratings of the teaching methods used in Physics 11b (spring 2007).

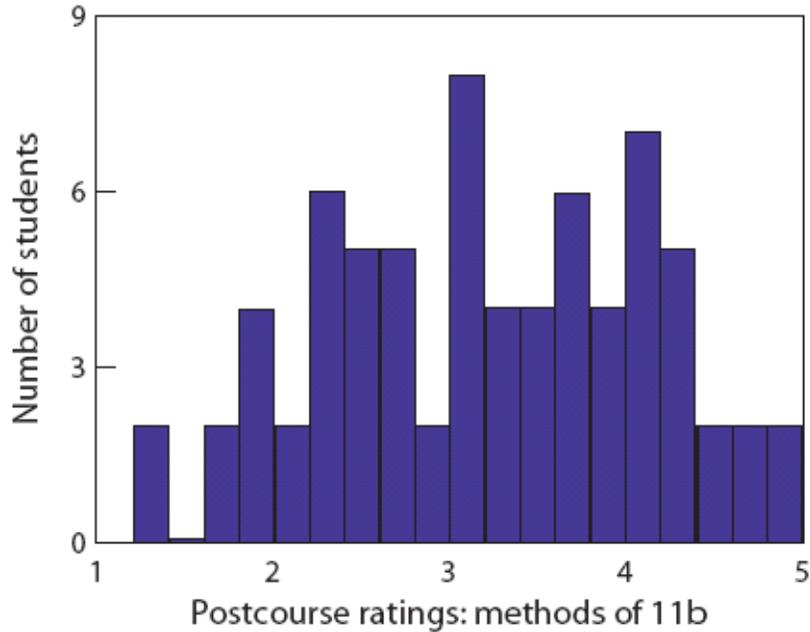


Table 3.8. Estimated bivariate correlation coefficients relating student ratings of the teaching methods used in Physics 11b and students' performance on conceptual tests and course achievement.

	<i>N</i>	<i>r</i>
FCI	74	0.23*
CSEM	90	0.22*
final exam	94	-0.05
final grade	94	-0.01

* $p < 0.05$

3.4.2 Relationship between student ratings and students' reported expectations about learning

In this section, we address our second research question, by examining the relationship between students' end-of-semester ratings on the interactive teaching methods used in Physics 11b and their responses to a survey about their expectations about learning. We estimated the bivariate correlation between *methods* ratings and scores on each of the clusters of items from the pre-course survey on student expectations (see *Measures* in the Methods section for the full descriptions of each cluster and associated questions). We present the estimated correlation coefficients and associated p -values in Table 3.9. Note that some students did not complete every question on the pre-course survey; for the analyses reported in Table 3.9 we removed any students who did not complete the entire survey.

As seen in Table 3.9, we find a statistically significant positive relationship between students' ratings of the interactive teaching methods and their self-reported expectations about the role of *conceptual understanding* when learning physics. Students who report that learning physics means achieving a deep understanding of the concepts tend to rate the teaching methods higher than those who report that physics is more about formula matching and memorization. This relationship is moderate in magnitude ($r = 0.41$, $p = 0.0006$), indicating that student responses to the pre-course survey questions on *conceptual understanding* explain about 16% of the variance in student ratings of the methods.

Table 3.9. Estimated bivariate correlation coefficients between students rating of the interactive teaching methods (*methods*) used in Physics 11b and their expectations about learning, based on their survey responses at the beginning of the course. ($N = 68$)

	<i>r</i>
concept understanding	0.41***
physics problems	0.15
role of professor	0.18
self-confidence	0.10
sense-making	0.22~
real world	0.21~

*** $p < 0.001$; ~ $p < 0.10$

These results indicate that students' reported expectations about *physics problems*, the *role of professor*, and their *self-confidence* are not statistically significantly related to their rating of interactive teaching methods. We do find a marginally statistically significant relationship between student ratings and their reported expectations about the role of *sense-making* and connections to the *real world*. Not surprisingly, the questions included in these composite variables are similar to those in *conceptual understanding*.

We regressed *conceptual understanding* onto *methods*, finding that for every one point difference in students' reported expectations about learning, there is an associated 0.53 point difference in their rating of the teaching methods.

3.4.3 Race and gender differences in student ratings of interactive teaching

In this section, we address our third and final research question, investigating race and gender differences in students' ratings of the methods used in Physics 11b. We also reconsider our first two research questions, in light of our results by gender. We summarize here the main findings before proceeding into the statistical details:

- On average, we do not find differences between races on responses to *methods*, our variable for measuring students' ratings of the interactive teaching methods;
- We do, however, find statistically significant differences in students' ratings, by gender, on average —female students in Physics 11b rated the interactive teaching methods lower, on average, than male students;
- Once we control for students' gender, the relationship between students' learning outcomes and ratings of the methods is no longer significant;
- Furthermore, once we re-examine students' reported pre-course expectations, we find that for male students the relationship between *conceptual understanding* and *methods* is positive and statistically significant, while for females the relationship is not statistically significant. These results suggest that other factors are contributing to female students' more negative ratings of the teaching methods.

We first examined whether under-represented racial- and ethnic-minority students rated the interactive teaching methods used in the course differently than their peers, on average. We did not find any statistically significant differences. Black and Hispanic students' average response to items about the teaching methods was 3.3 (SD = 0.8), while the class average was 3.2 (SD = 0.9). Similarly, we did not find any racial differences in students' overall ratings on the university administered student evaluations.

In contrast, we did find statistically significant gender differences in students' response to the teaching methods. In Table 3.10, we show the summary statistics for male and female students on student ratings of the teaching methods used in 11b, their self-reported expectations about learning physics, pre-course FCI scores, and measures of course learning and achievement (CSEM scores, final exam scores, and final grades). We present the number of students completing that assessment, as well as the mean and standard deviation for both genders. To compare the differences between males and females, we list the effect size¹⁴ (ES) for the difference and give the associated p -value, obtained either using a two-tailed t-test or Wilcoxon rank-sum test.

Table 3.10. Average differences in student ratings of the methods, their reported pre-course expectations about learning physics, their pre-course FCI and post-course CSEM scores, and other course achievement variables, by gender.

	female students			male students			difference	
	N	mean	SD	N	mean	SD	ES	p
methods	42	2.80	0.80	52	3.53	0.86	0.81	0.0001
concept_und	44	3.38	0.54	55	3.37	0.70	0.02	0.97
preFCI	47	22.70	5.30	55	25.87	4.14	0.64	0.0007
postCSEM	54	23.50	5.07	67	25.61	4.02	0.46	0.02
final exam	60	75.36	12.88	70	79.88	10.60	0.38	0.03
final grade	60	88.50	7.54	70	90.69	5.74	0.33	0.11

On the post-course survey about the course, the mean value for female students' rating of the methods is statistically significantly lower than male students' rating. This difference is almost $\frac{3}{4}$ of a point, on a scale from 1-5 and is statistically significant ($p = 0.0001$). At the beginning of the course, male and female students reported similar

¹⁴ Effect size is estimated by subtracting the female mean from the male mean and dividing by the pooled standard deviation.

expectations about learning physics, but we found statistically significant differences on student performance on the FCI ($p = 0.0007$). Male students obtained an average score more than 3 points higher than female students on the pre-course conceptual test on Newtonian mechanics, corresponding to an effect size of 0.64. On the two assessments at the end of the course, CSEM and final exam, male students outperformed female students, on average, but with a slightly smaller effect size. We found no statistically significant differences between the genders on final grade.

The effect of gender on student rating of the methods is dramatic; male students rated the *methods* 0.8 standard deviations higher than female students, on average. In the previous section, we found that students' pre-course FCI scores and post-course CSEM scores were correlated their end-of-semester ratings of the interactive methods used in Physics 11b. In Table 3.11, we again examine the relationship between students' performance on conceptual surveys and their ratings of the teaching methods, but now we also control for the effects of student gender. In Model 1, we control for student gender and examine the effect of students' pre-course FCI score on student rating. We find that when controlling for whether a student is male or not, students' pre-course FCI scores no longer have a statistically significant effect on their ratings of the course methods.

Similarly, controlling for student gender, we find that CSEM posttest scores are no longer statistically significant predictors of student ratings, as seen in Model 2. For our sample, student gender appears to have a dramatic effect on student ratings of the teaching method, even when controlling for learning outcomes.

Table 3.11. Parameter estimates from fitted regression models of gender and conceptual tests on student ratings of the methods used in Physics 11b.

	Model 1	Model 2
intercept	2.43***	2.21***
male	0.91***	0.70***
precourse FCI	0.009	
postcourse CSEM		0.02
<i>N</i>	73	88
<i>R</i> ²	0.26	0.19

*** $p < 0.001$

In Table 3.12, we present fitted regression models predicting student rating of interactive teaching methods by gender and student responses to the survey of their expectations about learning physics. We fitted these regression models in a dataset and corrected the analyses for missing data using the method of multiple imputations (Rubin, 1987) to obtain the parameter estimates. In addition to the variables included in the model, we used students' final exam and final grades to impute the missing data, to account for the achievement differences between students who had completed both the pre-course and post-course survey and students who were missing data one or both surveys.

Table 3.12. Multiple imputation estimates^a of parameters from fitted regression models predicting student rating of the teaching methods used in Physics 11b ($N = 130$).

	Model 1	Model 2	Model 3	Model 4
intercept	2.77***	0.90~	2.46**	2.94***
male	0.77***	0.81***	-1.28	-1.76**
concept_understand		0.55***	0.14	
maleXunderstand			0.62*	0.76***
upperclassmen			-0.29~	-0.31~

~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In Model 1, we confirm that male students rate the teaching methods higher, on average, than female students, by almost a full point on a scale from 1-5 ($p < 0.0001$). Model 2 includes a predictor describing *conceptual understanding*, a variable composited from student responses to items on the pre-course survey about student expectations. Students who reported believing that learning physics involves deep understanding of concepts tend to rate the teaching methods higher than students who reported expecting that physics learning is more about memorization and formula-matching ($p < 0.0001$). Controlling for student expectations does not reduce the effect of gender on student rating; the difference between male and female students' ratings remain statistically significant ($p < 0.0001$).

In the Model 3, we include the two-way interaction between student gender and their reported expectations about learning physics as a predictor. The interaction variable (maleXunderstand) has a marginally statistically significant effect on the outcome, suggesting that the relationship between student rating and their reported expectations is different for male and female students ($p < 0.05$). For male students, on average, every one point difference in *conceptual understanding* corresponds to a 3/4 of a point difference in student ratings of the methods used in the course. For female students, the relationship between students' reported expectations and student ratings is much lower. For every one point difference in *conceptual understanding*, there is an associated 0.14 point difference in female student ratings. We also include a variable distinguishing juniors and seniors (upperclassmen) from freshmen and sophomores (underclassmen).

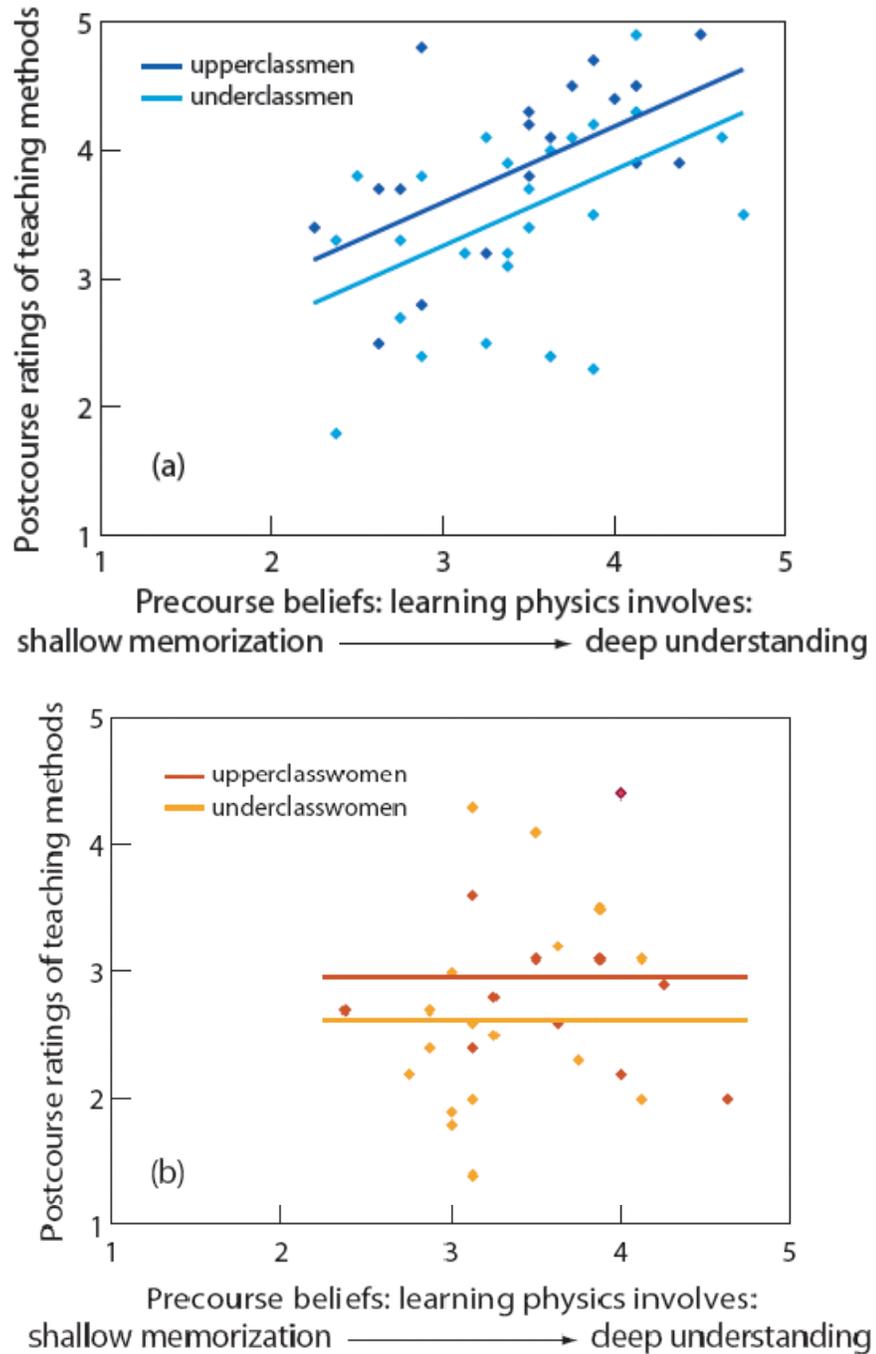
The upperclassmen rated the methods lower, on average, than underclassman, by 0.3 points.

In the Model 4, we remove the main effect of *conceptual understanding*; in effect removing the effect of female students' reported expectations about learning physics. Because the relationship between *conceptual understanding* and student rating of the methods is not statistically significant for females, we find that removing this variable improves our statistical power without affecting the shape of the regression model for male students.

In Figure 3.2, we display a scatterplot of student ratings of the teaching methods versus their reported pre-course expectations about learning physics as well as the fitted plots for prototypical individuals, using the final model in our analysis. Male students are presented in Figure 3.2a; female students in Figure 3.2b. Inspection of the scatterplot confirms the positive linear trend linking students' reported expectations and their ratings of the teaching methods for male students.

While there is a great deal of scatter among the data points for males, especially upperclassmen, there is a positive relationship between male students' pre-course expectations and their post-course ratings. Looking at the data points for females, however, reveals a drastically different picture. There does not appear to be any relationship between student reported expectations about learning and student ratings of the teaching methods, for females.

Figure 3.2. Scatterplots and prototypical fitted plots for students' reported expectations about learning versus their ratings of the interactive teaching methods used in Physics 11b. (a) male students ($N = 44$); (b) female students ($N = 29$).



3.5 Discussion

Our first research question asked about the relationship between student learning and student evaluations of teaching in a reformed introductory physics course. Our results were mixed and differed by level of course. In the lower-level introductory course, Physics 1b, students' scores on the end-of-semester conceptual survey had a statistically significant relationship with their evaluations of teaching: students with lower CSEM scores tended to rate the course lower. However, in Physics 11b, the more advanced introductory course, the strongest relationship exists between SET and their pre-course FCI scores. Students already entering the classroom with greater conceptual understanding of the prerequisite course tended to rate the course higher, suggesting that the incoming state of the students is more relevant for their evaluations of the course. Meanwhile, in Physics 11b the relationship between SET scores and CSEM scores was only marginally statistically significant. Looking across courses, the magnitude of this correlation is about 0.25, which is not very high and therefore only about 6% of the variability in student ratings is due to differences on CSEM scores. While this analysis suggests that there is a relationship between student learning and student evaluations or ratings of the teaching methods, the strength of this relationship is small in these reformed courses.

We next looked at the relationship between student expectations about learning and their ratings of the course. Using scores on the different components of the MPEX-2 and other expectation questions that we obtained from cluster analysis, we correlated students' pre-course survey responses with their post-course ratings of the teaching methods. We found a statistically significant relationship of students' expressed

expectations about the importance of deep conceptual understanding in this physics course with their rating of PI. Students who reported that they believe that to do well in this physics course they will need to understand concepts tended to rate the pedagogy higher than those who reported that they expected the physics course to involve more formula-matching and memorization. This correlation was moderate ($r = 0.41$), suggesting that the expectations of students play a valuable role in student evaluations of teaching, especially in reformed courses.

Our most dramatic finding was the gender differences in students' rating of the teaching methods. On a scale of 1-5, female students rated the methods about 2/3 of a point lower than male students. Female students also obtained lower scores on the pre-course FCI and post-course CSEM, although there were no statistically significant differences on their reported pre-course expectations. Furthermore, while there is a statistically significant relationship between male students' expectations of the course and their rating of the course methods, this relationship among female students is minimal and not statistically significant. In the next chapter, we examine reports from student interviews to better explore why female students rate PI and the interactive teaching methods lower in this course.

As with the rest of this dissertation, the main limitation to this research is the unique student population that we are studying. Furthermore, the response rates to our surveys were low, so we are only sampling the subpopulation of students willing to report their expectations and ratings of the course. However, despite these limitations, these results do point to the need for better understanding how students, particularly female students, evaluate courses taught using reformed pedagogies. Following up on these

results, we narrowed our analysis in the next chapter to specifically address how female students responded to this course.

Chapter 4

Understanding female students' responses to interactive teaching

4.1 Introduction

In their review of gender and science education, Brotman and Moore (2008) describe several features of gender-inclusive pedagogy, including drawing upon the experiences, interests, and preconceptions of both genders; prioritizing active participation; emphasizing collaboration and communication; and providing a supportive environment. Many of these features can be found in research-based reforms in introductory physics courses at the post-secondary level, including the reforms described in the previous chapters: Tutorials in Introductory Physics (McDermott, et al., 2002), JiTT (Novak, Gavrin, et al., 1999), and PI (Mazur, 1997). In light of these similarities between Brotman and Moore's gender-inclusive pedagogy and the reformed teaching methods used in introductory physics courses at Harvard University, our findings from the last chapter are surprising: female students rated the teaching methods used in their introductory course lower than their male peers. In this chapter, we examine student responses during interviews in this same course to consider why female students responded to these methods, specifically PI, negatively.

4.2 Background

In this section, we review relevant literature that describes gender differences in participation in, and response to, small-group learning. We focus on the few studies at the college level that have described negative female experiences in interactive classrooms. We also examine research at the secondary level, specifically from small-group learning in science classes, which provides insight into how and why female students may experience these interactive pedagogies differently from male students.

4.2.1 Stressful collaborations

Many of the reforms in introductory physics courses involve students interacting with each other as they engage with the material, however research suggests that these interactions may be stressful to female students. Workshop Physics (Laws, 1991) is a research-based method for teaching calculus-based introductory physics that eliminates formal lectures and uses collaborative activities in which students make observations, analyze data, and develop verbal and mathematical models. Workshop Physics has been immensely successful in improving gains on conceptual surveys and is one of the few reformed pedagogies to obtain gains on epistemological surveys in introductory physics courses (as described in Redish, 2003). Although these teaching methods offer many benefits for student outcomes, Laws, Roosborough and Poodry (1999), in research conducted at a private liberal arts college in the northeastern region of the United States, found that junior and senior women responded more negatively to these teaching methods

in introductory physics courses,. In a survey of students' attitudes toward different learning experiences (e.g. textbook reading, using the computer), junior and senior female students reported more negative attitudes about laboratory activities after one semester of Workshop Physics. Furthermore, the upperclasswomen's end-of-semester responses were more negative than their male and underclasswomen peers.

In follow-up focus groups with current and former female students, Laws et al. (1999) found that "stressful collaborations" were often the cause of frustrations of the female students. "Women complained of domineering partners, clashes in temperament, being subjected to ridicule, fears that their partners didn't respect them, and feelings that their partners understood far more than they." (p. S35) Furthermore, many of the other concerns that students reported are attributed to issues with their beliefs about learning, specifically viewing learning as obtaining the right answers. "The results of the 1990 attitude survey and our impressions in other years lead us to believe that women, more than men, have been encouraged to view learning as straightforward fact gathering or memorization." (p. S35)

Although the sample size for the survey was small ($N = 44$; $N_{\text{upperclasswomen}} = 9$) and the qualitative methods and analysis were informal, these results from Workshop Physics at Dickenson College suggest that male and female students may indeed respond differently to specific implementations of small-group inquiry-based teaching methods. Specifically, females may report more negative responses to the social issues surrounding the in-class collaborative work and the conflict between the course goals and students' ideas about learning physics.

4.2.2 Gender differences in discourse

Research on student discourse in elementary and secondary science classrooms also offer evidence on gender differences in small-group collaborations and the social relationships therein. Jones, et al. (2000) studied dyads of students to examine patterns of language and tool use in urban elementary science classes in southeastern region of the United States. Through analysis of video and transcripts of student collaborations, the authors coded for when student speech was individual-centered, dyad-centered, or commands. The authors first observed that the male students' discourse was primarily individual-centered, or commands. Furthermore, they found gender differences in the dyad-centered speech. Female students were more relational in their approaches to each other, exhibiting a sense of caring for their learning environment and partners. The authors present female student discourse showing students attempting to maintain their relationship in addition to completing the tasks. In contrast, in addition to occurring less frequently, much of the male dyad-centered discourse was less focused on preserving relationships, instead focusing on competition or teasing. While the magnitude of these gender differences is unclear from the evidence the authors present, the qualitative results provide further evidence of how social relationships may play a role in how female students experience collaborative work in science courses.

While the previous work is suggestive of gender differences in student interactions in small-group work, more convincing evidence is presented in Alexopoulou and Driver's (1997) study of 14-15 year-old Greek students in small groups discussing qualitative physics questions. Researchers coded utterances involved in the social interaction (e.g., agreement, disagreement, question, extension) and argument

construction (e.g., making a prediction, asking or providing evidence for an assertion). Labeling discussions as “progressive” when one or more members of a group “progressed” in their reasoning, Alexopoulou and Driver found that between male students there were more disagreements and asking for evidence, while female students tended to agree with each other, but then ask questions. Furthermore, in “regressive” discussions, they found that groups of females appeared more concerned with maintaining consensus rather than exploring their ideas, often agreeing on the most common response. These gender differences in discourse patterns were statistically significant and offer the most concrete evidence that males and females interact differently in small-group discussions about physics.

4.2.3 Inequity in participation

While gender differences in discourse suggest that male and female students may experience collaborative work differently, other gender differences in participation were found in introductory college physics courses using Cooperative Group Problem-Solving at a large public research university in the midwestern region of the United States (Heller & Hollabaugh, 1992; Heller, et al., 1992). Researchers found that small groups consisting of two male students and one female student did not perform as well as homogeneous or female-majority groups (Heller & Hollabaugh, 1992). In observations of these groups, researchers found that the sole female student’s contributions to the discussions were not valued, regardless of her physics knowledge or articulacy.

“For example, during their work on a projectile motion problem, a group with lower-ability male, a medium-ability male, and a higher-ability female had a vigorous discussion concerning the path a projectile would follow. The men

insisted on a path following the hypotenuse of a right triangle, while the woman argued for the correct parabolic trajectory. At one point she threw a pencil horizontally, firmly commenting as it fell to the floor, ‘There, see how it goes—it does *not* travel in a straight line!’ Even so, she could not convince the two men, who politely ignored her arguments.” (p. 641)

Similarly, in secondary mathematics classes, Webb (1984) found that in male-majority groups the male students tended to ignore their female group members and direct their explanations solely to other males. Often, a female student was ignored even when explicitly asking for help, if she were working in a group with three male students.

These observations align with the literature on gender and social influence and status. In a review of research on gender differences in social influence, Carli (2001) found that males exert more influence than women in mixed-sex groups, particularly in stereotypically-male contexts. These gender differences in influence are in part attributed to differences in perceived competence; to influence others one must demonstrate competence, however often females are perceived as less competent than their male peers, even when there are no objective gender differences in performance or behaviors. Furthermore, Carli (2001) also asserts that a female’s competent and assertive speech may be perceived as threatening to others, particularly males, and therefore reduce her influence. While this conflict between displaying competence but not in an assertive manner can inhibit females’ persuasiveness, demonstrating friendliness, warmth, and other communal characteristics can help. “A man can influence others even when they do not particularly like him, but a woman must be likable to be influential.” (p.734)

Similarly, issues of status can affect students’ participation in small groups. In a study of Complex Instruction (E. G. Cohen, Lotan, Scarloss, & Arellano, 1999) in an ethnically diverse, urban middle school science class in the western region of the United

States, Bianchini (1997) found that students with high status—students who are perceived as smart and popular by their peers—participate more in discussion, have greater access to materials, and achieve greater learning gains. In contrast, “low-status students were often seen as unable to provide intellectual insight, rarely asked to voice their opinions, and allowed to do little substantive work.” (p. 1055) Similar results were found in other contexts (e.g., E. G. Cohen, 1984; E. G. Cohen & Lotan, 1995) and although gender was not addressed explicitly, it is often given as an example of status characteristics.

In her review on socio-cultural approaches to mathematics cooperative learning, Esmonde (2009) noted that it is not students’ perceptions of each other that can impact their participation, but also students’ concerns about how they might be perceived by others, specifically for stereotyped groups. “[With] the activation of stereotype threat... African American students or girls might participate less in mathematical discussions, not because they believe themselves to be less competent, but because they expect to be evaluated in ways that reinforce undesirable stereotypes.” (p. 1022) The impact of stereotype threat has been well shown on standardized tests (Steele & Aronson, 1995); the risk of being judged by a negative stereotype depresses the performance of the stereotyped group, such as females on a math exam (Spencer, et al., 1999). Thus, how students perceive stereotypes about women and physics may be an important part of the context female students contend with in small-group discussions in their introductory physics courses.

4.3 Methods

4.3.1 Interview participants

We conducted this study in the same Physics 11b course described in the previous chapter. This course was the second-semester course in a one-year sequence in introductory physics. Requiring at least concurrent enrollment in multivariable calculus, Physics 11b covered topics in electricity and magnetism, circuits, and optics using a draft research-based textbook written by the course instructor. The course consisted primarily of a mix of biological science and engineering science majors and was split evenly between males and females (54% male).

A variety of interactive teaching methods were used in the course, with an instructor experienced in using these methods. To help students prepare for class, the instructor used JiTT: before class, students were required to read the textbook and answer questions about the material. During the large lecture meetings, the instructor used PI (Mazur, 1997), first reviewing the material by giving short presentations and then asking several ConcepTests during the 1.5 hour meetings. Students received participation credit for responding to these questions in-class, but did not receive extra credit for answering correctly. The course also included two-hour optional sections run by graduate teaching assistants each week. During the first half of the section, students worked in small groups on Tutorials in Introductory Physics (McDermott, et al., 2002). During the second hour, the teaching assistants led cooperative problem-solving sessions, to give students practice with quantitative problems and help them with their homework. A more detailed description of the course methods can be found in chapter 2.

Before the semester, we recruited interview participants using the course list from the first semester course (Physics 11a). We offered small monetary compensation to students for each interview in which they participated. Ten students responded to our email request, each of whom agreed to participate. One male student decided to not participate in future interviews after his first interview. After the course began, we discovered that a number of students from the more advanced level first-semester course (Physics 15a) dropped down to Physics 11b. We emailed this sub-population of students to ask if they would like to participate in our exit interviews; two students (one male and one female) agreed to participate. Details of the eleven students who completed the interviews are presented in Table 4.1. Note that the students who participated in the course are not representative of the course enrollment; almost all interviewed students obtained grades of A- or higher and performed well on the conceptual surveys. However, these students do report a variety of overall ratings, ranging from 2–5 on a 5-point Likert scale. Note that only four out of the eleven students interviewed were female; at the start of the study we were not looking for gender differences, therefore we did not target our sampling (or interviews) for this purpose.

4.3.2 Interview data

We interviewed students at three different points in the semester: before the first class,¹⁵ in the middle of the semester, and at the end of the course. During the pre-course interview, we asked students about their experiences in the fall semester introductory physics course, specifically which features of the course they felt best helped them learn the material. We also asked whether they had heard about the teaching methods that

¹⁵ One student, V1, was interviewed after the first class due to scheduling conflicts.

would be used in Physics 11b and to describe how their ideal physics course would be taught to help them best learn. The second part of the pre-course interview focused on validating several of the pre-course survey questions, followed by a few content questions designed to more indirectly probe students' ideas about learning introductory physics. All interview protocols can be found in Appendix 3.

Table 4.1. Background information on interviewees.

	gender	first semester	FCI	CSEMpost	final grade	overall rating
A1	M	11b	29	30	A-	3
Y1	M	11b	29	27	A-	3
S1	F	11b	17	22	B-	3
G1	M	11b	29	30	A	5
K1	F	11b	28	29	A	2
V1	F	11b	22	25	A-	2
G2	M	11b	29	30	A-	4
M1	M	11b	29	29	A	2
N1	M	11b	20	20	A-	3
U1	F	15b	-	30	A-	2
D1	M	15b	30	25	A-	4

In the mid-course and post-course interviews, we asked students several questions that asked what they did and did not like about the teaching methods used in Physics 11b. These interviews were semi-structured; the interviewer asked all students the same set of questions, but asked follow-up questions based on the student responses. We also asked questions of students in order to refine questions on our post-course survey. During the mid-course interviews we also asked students to think aloud, as they solved a conceptual exam problem from a previous year.

All interviews were conducted one-on-one and students were notified that none of the teaching staff would know who participated in these interviews. Furthermore, student

comments were not shared with the instructor until after the course was complete, unless requested by the student to give the instructor formative feedback about the course. As mentioned above, eleven students participated in the mid-course and post-course interviews, seven of whom participated in both, for a total of 18 interviews about the course and the teaching methods used.

4.3.3 Analytic methods

In these interviews, we looked for patterns in the data that might reveal how female and male students experienced and/or responded differently in a PI environment. We first read through each interview several times, to obtain a general sense of students' responses to the course. We then categorized the positive and negative comments that students made about the course, developing general themes about what students liked and disliked about the course.

To explore gender differences in student responses to these methods, our analysis was both data-driven and theory-driven. We compared males' and females' positive and negative comments about different aspects of the course, examining differences in the number of students as well as number of comments. We also used the background literature to guide our analysis, focusing on issues in which gender has previously been a factor in group learning. From these analyses, we develop propositions to shed light on the dramatic differences in males' and females' ratings of the interactive teaching methods in this course.

4.4 Results

In our mid-course and exit interviews, students remarked on what they liked and disliked about the course, including their experiences learning in a PI environment. In this section we focus on two categories of student comments, *social issues* and *testing*, which suggest that male and female students may participate, experience, and respond to PI differently. These themes were present in many of the studies we discussed in the background section and, for *social issues*, females disproportionately mentioned these concerns in our interviews. Using students' interview data, we describe these two themes and provide evidence for how these issues may translate to different experiences with, and therefore different perceptions of, the pedagogical techniques used in this course. In Table 4.2, we provide brief descriptions of the themes and subthemes discussed in this section.

Table 4.2. Social issues and testing themes that emerged from our interviews.

Social issues contribute to differing experiences for males and females in this course

Discussing ConcepTests often involves navigating social relationships.

Friendships play an important role in a PI class.

Students are aware of social norms and atmosphere during discussion

Contrasting case: male comments about discussion without mention of social issues

Social issues linked to limiting the participation of female students during discussion

Some male and female students perceive a “testing” atmosphere during PI and respond negatively to the stress of answering ConcepTests correctly

4.4.1 Social issues contribute to differing experiences for males and females in this course

In our interviews, we asked students their thoughts about the use of ConcepTests and discussions with their peers in their introductory physics course, specifically whether they felt these activities helped them learn. In response, female students spontaneously mentioned *social issues* as a concern or factor in their experiences with Peer Instruction in all six interviews. In contrast, only one male interviewee mentioned social issues in the context of PI in the twelve male interviews we conducted. In this section, we describe the social issues (e.g., navigating the teaching and social relationships, needing or relying on friendships, and overcoming the social atmosphere) that the female students (and one male student) report. We also discuss how these issues may have shaped these students' experiences with PI and, in one student's case, negatively affected her perceptions of the pedagogy.

Social issue #1: Discussing ConcepTests often involves navigating social relationships.

Several comments by female students in these interviews suggest an awareness of how PI affects social relationships between students in the course, specifically during peer discussions. K1, a female student from Physics 11a, was especially attuned to how the discussion part of PI often involved one student teaching another. She made comments suggesting that this teaching relationship conflicts with the friendships and typical social relationships between students in the course:

...it's hard to make friends in a teaching environment when one person is kind of teaching the other. So – but I guess after the class, that relation – the teaching relationship can kind of end, and all the awkwardness will be gone. [K1 postcourse]

K1 describes her experiences in the PI classroom as “one person... teaching the other,” as opposed to students working together to learn the material. Therefore, students must manage the power shifts in their relationships when one student is the teacher, or authority figure, while they teach their fellow students. K1 explicitly refers to the navigation as “awkward.” She states that is hard to make friends—a relationship that implies equality—with another student when there is this teaching relationship as found in the PI environment.

In her mid-course interview, K1 elaborates on her thoughts about why a teaching relationship among students may be problematic:

It takes a very special person to be able to teach someone else and not to... have them feel like they're being taught really – or condescended to. That's really difficult and not all of us are so good at that. [K1 midcourse]

Here K1 emphasizes the importance of the ability to “teach someone else” in a way that reduces the awkwardness of being “taught” by your peers during PI. The students who can minimize their role as the authority figure in the discussions, even while they are teaching, will help their peers not feel “condescended to.” Unfortunately, K1 indicates that this skill is not easily found in the PI environment.

K1 was not the only female student to feel that that to teach your peers you need a “special” personality. V1 re-iterates this sentiment in her midcourse interview when asked about whether the discussion would be more helpful if students around her knew the answers to the ConcepTests:

I guess it would be really helpful if people around me knew a lot more than me, and were good at explaining things, and patient enough to explain things... I mean I guess it really depends on the personality of the person. Because, yeah,

because people are very good at explaining, and very patient, and don't make you feel like you're dumb, [laughs] but other people, not so much. [V1 midcourse]

In contrast to K1, V1 suggests that she is more favorable to peer-teaching relationships. She states that if another student taught, or explained things, to her, it would be helpful. However, she also talks about how a “teacher” can make you feel “dumb,” much like K1 describes not wanting to feel “condescended to.” V1 highlights the importance of patience and interpersonal skills in when students are teaching each other, but again suggests that these attributes are not always found in her peers.

In summary, female students that we interviewed described how the discussions during PI can produce teacher-student relationships among peers. Furthermore, they discuss how this change in their interactions with their peers can produce “awkwardness” and make them feel “dumb,” especially when the peer teachers do not have the interpersonal skills to navigate the different relationships in a PI classroom. While attributes such as patience and good communication are important for PI, all of the female students and one male student also discuss the importance of pre-existing friendships and relationships. We describe this theme in the next section.

Social issue #2: Friendships play an important role in a PI class.

In the previous section, we described how some of the female students felt that they needed to navigate the different relationships with their peers during the discussion, *e.g.* social relationships and the teaching relationship that may develop. However, navigating between these relationships may be easier when students are already good friends prior to entering the course, as K1 described in response to an interviewer's follow-up question about why it would be hard to be in the course without friends:

...if I didn't understand a question, or if I was just having a day where I didn't want someone to, like, try and explain something, like, my friends would pick up on that, and they wouldn't push it. And they're more sensitive to how you learn and what your likes and dislikes are and what will offend you and what won't offend you. With strangers, you know, there's more room for error. And I think if I didn't have any friends in a course, I probably wouldn't have discussed very much just for fear of having those conflicts with people I discussed with. [K1 postcourse]

K1 talked about how her friends could be more sensitive to her feelings and preferences when discussing the ConcepTests. She described how the peer discussion does not only involve the physics concepts, but that students also have to be attuned to the affective state of their peers, *e.g.* when they are having a bad day or when they might be offended. Furthermore, K1 stated that she would have reduced her participation without these pre-existing friendships.

We interviewed another female student who was just in the situation K1 described. During the interview, S1 stated that only knew one person in the class and only sat by that person. Later she went on to say that the course did not do very well to develop collaborative problem-solving skills because “it’s the nature of people to sit by people they know... you can’t really, like, force people to talk during class.” Without pre-established friendships, S1 claimed that students were less willing to collaborate and discuss the ConcepTests with each other.

While other students we interviewed did not mention not having enough friends or acquaintances in the course, one female student (U1) and one male student (D1) did mention that the discussion may have been improved if they sat with different people.

U1: Yeah, because sometimes I feel like discussing with a neighbor, we kind of go around in circles for a long time, because we all don't know what's going on. Or

a lot of times it's either we all know what's going on or we all don't really know what's going on.

Interviewer: So do you think it would have been, um, a different experience if you were sitting next to people that had a different background than you?

U1: ...Um, I do think so, yeah. But it's just that you kind of tend to want to sit with your friends, yeah [laughter]. It just never happens that way. (U1 postcourse)

U1 started off describing some of the unproductive discussions she had with her neighbors. When asked by the interviewer about whether her experiences would have been more positive had she sat next to different people, U1 responded favorably. However, she went on to state that although she believed it would help her, she preferred sitting with her friends. Again, friendships appeared to mitigate student experiences with PI. Similarly, D1, a male student from Physics 15a, also talked about the role friendships play in student experiences with PI.

Um, the - the biggest part about the ConcepTests for me was the discussion period and disagreeing with someone, which makes you defend your position, um, although there were quite a few times where everybody would just be, like, "It's 1, it's 1, it's 1," you know, and everybody would agree, everyone around me. You know, "What do you get, 1?" Yeah, [inaudible], but that doesn't really do anything for anyone. Um, and - and I - I guess, uh, the professor would say, like, "Sit around people that, um, you don't - or that don't get the same answers as you," but generally, uh, that's - that's really hard to do because, uh, the people, uh, well, like, the people at least that I know tended to get the same questions right and the same questions wrong. Um, and I think it'd be hard to find someone who disagreed with you on a consistent basis - um, unless you're, like, you know, in a large group, then there's sure to be some disagreement in quad. Uh, and also, it's hard to sit away from your friends. (D1 postcourse)

D1 also indicated that friendships were influential in his experiences in this course. He acknowledged that discussion would be better if the people with whom they were talking had different answers or experiences, but suggested that he was more likely to sit with

their friends even if they all tended to have the same answers or ideas about the ConcepTests. Interestingly, these comments by D1 are the only male comments in our interviews about social issues in PI course. Although he says “it’s hard to sit away from your friends,” he also talks about how he appreciates disagreeing with other students and “defending your position” in the discussions. His comments offer contrast to those from the previous section, as he talks about these peer interactions in detail, without mentioning the social issues that may arise from disagreements.

All female students and one male student suggested that with no or few friends, or friends with similar background knowledge and ideas, PI discussions are more limited in their value. Furthermore, lack of friends or acquaintances directly impacted the experiences of one female student in our sample, who only sits by one other student throughout the course. Therefore, pre-established friendships and relationships appeared to impact students’ experiences in this course, especially for female students.

Social issue #3: Students are aware of social norms and atmosphere during discussion

Asking students to talk to one another during lecture also means asking students to change their ideas about what should be going on during class. In a traditional lecture environment, students typically sit down, open their notebook, and listen and copy down while the instructor presents the material. In a PI classroom, all students are expected to actively participate by answering questions and discussing with their neighbors. Participating in PI classroom requires changing behavioral expectations or norms in the classroom. One female student, V1, talked about this change in her pre-course interview, which, for logistical reasons, took place just after the first class. When asked what she

thought about the teaching methods, specifically PI, she talked about how she didn't think the class would be that willing to participate:

V1: I don't really think people are gonna get as into it as he, like, you know, he has mentioned, like, climbing over chairs and – to find someone who disagrees, but um, I, I doubt that it's going to be that, you know –

Interviewer: Why not?

V1: Um, I feel like people don't get that passionate about things in class, like, um, even in, like, you know, like, when you talk about ethics, and, like, things that, you know, people should be passionate about. A lot of people aren't in, like, a lecture setting isn't really a place where people, like, get all riled up and, like –

Interviewer: Why do you think that is? What is it about the lecture setting?

V1: Um, I don't know. I, I think people just have this idea of, like, lecture, you know, you can sleep in lecture, or, like, you can, you know, lectures are optional and things like that. Um, I think part of it's, like, like, you know, you don't wanna be, you know, dorky, and, like, too into things. You know? Um, but, I, I don't really know. (V1 precourse)

V1 comments were not necessarily about her own feelings about teaching and student participation, but more about her ideas of how other people respond and behave in a lecture setting: “a lecture setting isn't really a place where people, like, get riled up.” While she could be deflecting her personal feelings by projecting them on other people, her last comments suggested otherwise. When asked why people tend to not get passionate during lecture, she talked about not wanting to appear “dorky” or “too into things.” V1 predicted that the social norms of how one should participate (or not) in lecture will constrain how “into it” students get during discussion.

In her mid-course and post-course interviews, V1 again discussed the social atmosphere, confirming her prediction that students will not “get that passionate” when discussing the ConcepTests:

Interviewer: So, um, what about in lecture? So you have these ConcepTests. You have these questions. What do they do for you? Do they do anything? Like do they help you learn at all?

V1: ...I guess it's - I mean, I think they want much more vivid discussion in between the first answer and the second answer, which, at least where I sit towards the back, it doesn't really - it's not very animated, nor is it - you know, it's like, what did you put? Two. Me too. Okay, good. You know, like, it's not... they want us to jump chairs, and argue with someone. I don't think that's - there's not that much intensity [laughs] around where I sit. (V1 midcourse)

When asked whether ConcepTests help her learn, V1 again focuses on the social atmosphere related to the discussion. V1 refers to her seat location “in the back” and talks about what she and her peers do and say during the discussion, claiming that “there’s not that much intensity around where I sit.” These comments, in light of her comments during the pre-course interview concerning not wanting to appear “too into things,” suggest that the social norms and atmosphere may have prevented V1 from participating more actively in the discussions.

Contrasting case: male comments about discussion without mention of social issues

A1 is a male student who took Physics 11a and also comments at length about the discussions. He finds that the discussions of ConcepTests are most helpful when he is confused, and, similar to U1 and D1, that the other students he sits with often either all have the same answer or are all confused. When asked about whether the quality of discussion would change if he sat next to different people, A1 talks about how it would be hard to discuss with students with different learning styles:

Interviewer: Do you think [the quality of discussion] would change if you sat next to different people?

A1: Um, it would depend, I guess, on what their learning styles were like. Um, as it is, I like to learn things quickly. So, um, for example, when we don't understand something, I feel like the group I'm in, we can debate it, like in the same style, I guess. That may not make a lot of sense, but, um, if there was someone who had a different learning style, it might be harder to debate questions... I mean, so I'm – I guess I like to use the equations if it were mathematical, whereas some people like to reason things out, and just how we would convince other people of our answer would be different depending on which style you prefer, I guess. So like if I used the equation, you know, that potential is related to the charge – if the charge goes down, the potential has to go up. But, um, if other people do it conceptually, then you have to explain it differently... (A1 midcourse)

Although A1 discussed at length issues related to talking about the ConcepTests with his peers, he did not mention the social issues that all female students (and one male student) brought up: the need to navigate social relationships, importance of friendships, and influence of social norms and atmosphere. In contrast, A1 talked about how students approach learning differently and use different justifications for their answers to ConcepTests. Although these differences could lead to social issues, he did not spontaneously bring these up in our interview. Note that A1's ideas were not representative of the other male students we interviewed; many had different ideas about discussion, including that the discussions were unnecessary for their learning, but similar to A1, none of them mentioned the social issues brought up by the female students, aside from the one comment by D1.

Social issues linked to limiting the participation of female students during discussion

All of the female students we interviewed made some mention of social issues that may arise during Peer Instruction. Furthermore, their comments suggest that managing different peer relationships and dealing with the social norms and atmosphere changed or limited their participation in peer discussions: S1 reported only discussing the ConcepTests with the one person she knew in the course, U1 and D1 report sitting only

with their friends even though they tended to all have the same answer, and V1 suggested that she was influenced by the social atmosphere and norms near where she sat and therefore did not actively participate in the discussions. K1, who seemed especially attuned to the teaching relationships that may result during peer discussions, said “I generally don’t take part in those discussions too much because I do feel like I just don't wanna make someone feel like I'm trying to tell them what to do.” (K1 midcourse) Each of the female students and D1 presented evidence that their experiences with PI were impacted by the social issues they encountered.

The greater attention to the social issues appears to have contributed to different experiences in a PI classroom for the female students than for the male students we interviewed in this study. However, it is unclear from these interviews whether the social issues contributed to females’ poorer ratings of PI. There is evidence from one female student, K1 who reported not only limiting her participation in the discussions, but responding more negatively to the teaching method due to the social issues she experienced during discussions:

Interviewer: So you have these questions in class, you know, the - where the professor asks you a question... do they help you learn at all?

K1: Oh, um, it's hard. I feel like if I get the answer right, I'm learning; but then when I - I know this isn't, like, what he was hoping to have, but when I get it wrong, then everybody turns to you, and, like, all your friends are, like, trying to tell you what's going on. And I just don't do well with everybody telling me - and one time the girl in front of me - and I know we're supposed to talk to people we don't know, but I think it's even harder to learn from people you don't know - it's hard to take criticism from people you do know, but then just to have - this girl whips around in her chair and says, "No, that's not the answer," and she started, like, yelling at me. I was just not prepared for that...

Interviewer: So you mentioned why, um, that you don't think you learn, like, people in general don't learn from people that you - they don't know very well. Why - why is that?

K1: I think it's a - kind of an interpersonal thing. You have more of a tendency to feel attacked... (K1 midcourse)

In response to whether the ConcepTests help her learn, K1 brought up her negative experiences discussing the questions with her neighbors. She viewed someone disagreeing with her answer as “criticism” and felt “attacked” when discussing with people she doesn’t know very well. She barely talked about the questions themselves, only saying briefly that she felt like she was learning when she got the correct answer. Her negative social experiences during the discussion seemed to dominate her opinions on the usefulness of the ConcepTests and pedagogy in general. While K1 was the only female student we interviewed who clearly linked social issues to her negative opinion of the pedagogy, her comments demonstrate how these issues could negatively impact female students’ experiences in an interactively-taught course.

4.4.2 Some male and female students perceive a “testing” atmosphere during PI and respond negatively to the stress of answering ConcepTests correctly

In addition to dealing the social issues related to the discussion, we also found that students viewed the in-class questions as ‘tests’ as opposed to opportunities for learning. During one of the midcourse interviews, a male student remarked that he did not like the ‘testing’ that came with answering and discussing the conceptual questions in class:

Interviewer: Um, so you're in class and you have these questions, and on these ConcepTests you say you use these clickers. So what do you think they do for you in terms of your learning? Do you think they help at all?

N1: Personally, I find them very stressful. The course relies very much on testing. And that's probably also the only negative I can say...sometimes it really feels like you're over-saturated with test things. You get like several questions - - um, during - during the lecture, and you have homework, you have lab, um, all of course is graded. And then you have sometimes pre - pre-surveys and they are graded. And the readings are graded. Basically, everything you do is graded, and - and it's a lot, and you're just like, "Wow! You know, I need a break here." So at - at some point I was like, "This is too much." I don't like the testing. (N1 midcourse)

His comments are surprising, given that in this course students do not receive credit for right answers but for completion when answering the ConcepTests (same with the pre-surveys and reading assignments). N1 indicates later in the interview that he realizes that he is not being graded for correctness, yet still feels as though he is being "tested" when responding to the ConcepTests:

Interviewer: Do you feel pressure even though you know you're not being graded on correctness?

N1: Um, yeah. I think, I mean, maybe that's just the [inaudible] because you always try to compete, and you always want to perform your best, so that's probably the pressure that you also create for yourself. I'm pretty sure that most of the people here are quite motivated to do well. And um, reinforcing that might go overboard, case in point, and take it a bit too far. I-I-I didn't mind the questions, I just, you know, wanna do 10 to 12 questions sometimes. It's too much. (N1 midcourse)

N1 states that he feels pressure to get the correct answer, which he equates to "competing" and "performing his best," even though whether he answers the ConcepTests correctly does not impact his grade. He also references being at a highly selective institution by making the point that "most of the people here are quite motivated

to do well.” Interestingly, he views the ConcepTests as exacerbating the pressure on students to perform.

N1 is not the only student to talk about ConcepTests as “tests.” After hearing his remarks, we asked the other interviewees whether they felt that they were being ‘tested’ with the interactive teaching methods. Out of the fourteen students we interviewed, half remarked on the pressure to get the correct answer or feeling as though they were being “tested.” As a consequence of this pressure, students felt that it would be a negative experience if you don’t answer enough ConcepTests correctly. For example, another male student hypothesized how a student would feel in the course when getting too many of the questions wrong:

D1: I suppose it would be difficult for a student who, uh, doesn't get a lot of them right and is sitting next to someone who does because, uh, well, that's just kind of demoralizing - um, sort of a difficult thing to avoid, though. Um, but, uh, I mean, if you - if you did the reading and you think, okay, I understand this, and then you get to class and you had some ConcepTest that everyone around you seems to get it, but you don't, uh, I can see that as definitely leading to a negative experience. Um, I don't know how to avoid that, haven't thought about that.

Interviewer: Okay. But you didn't feel that way in general in the class, aside from midterms?

D1: In general, everyone I sat next to got about the same number right. Uh, so maybe that's a problem with my thought of having more difficult ConcepTests. Um, I think I'm fine if I only get maybe 50 percent of them right. But it might be kind of a struggle for people who would only get, like, 25 percent. (D1 postcourse)

D1 sympathizes with the ‘struggle’ he thinks students with too few correct answers will face, although he himself does not report experiencing this strain. Interestingly, his comments also suggest that it’s not only important to get a reasonable percentage of the ConcepTests correct, but it’s also important to get similar percentages

of correct answers as those around you: “I suppose it would be difficult for a student who doesn’t get a lot of them right and is sitting next to someone who does...” Again, D1 does not state that he feels the effects of this personally, as those around him in class get similar numbers of questions right, however he draws attention to the intersection of the social issues brought up in the previous section with the ‘testing’ atmosphere. K1, a student who expressed a great deal of concern for the social issues related to PI, also comments on the implications of the ‘tests’ on students’ social interactions:

The only thing that I still - like, I already feel like people compare grades enough here and say, like, “How did you do on the exam?” And this sometimes felt similar, like, “Well, what did you answer?” And it becomes uncomfortable [inaudible]. And everyone answered No. 1, and one person has No. 3...

I think what - I’m not really opposed to being tested, but I am kind of opposed to being tested in front of all your peers, which is kind of what the questions did, you know, in asking, “What did you answer? What did you answer?” (K1 midcourse)

Again K1 mentions the specific context of a highly selective institution: “I already feel that students compare grades enough here.” She likens the ConcepTest answers to comparison of test scores, framing the questions as opportunities to compare, and perhaps compete, with her peers. She refers less to the internal pressure brought up by N1, but the pressure to perform in front of her peers, similar to the comments by D1. She describes this experience as “uncomfortable,” again sympathizing with the student who has an answer different from all of their peers. Other students describe similar, or more negative, emotions when talking about the testing atmosphere, using words such as “frustrating” and “stressful.” In the quote below, V1 elaborates on why students feel anxiety when answering these questions in class:

Interviewer: Okay, so you're sitting in class, and the professor poses a question, and immediately you don't know the answer. How do you feel? Like what's your typical feeling or emotion that happens when he's posing these questions?

V1: Usually it's like, man, am I supposed to know this? And then I usually take out my textbook and try to re-read parts, but, yeah. Mostly - and then, mostly we turn to each other, and we're like, "Man, do you know what he's talking about?" [Laughs] That kind of thing. I just feel like, um, I guess, anxious to see whether it's like - 'cause I can't really tell if it's something we're supposed to have learned in the reading and I just didn't learn it, or if he's introducing something new. Like in regular lectures where, like, new concepts come up in class, so - So it makes you feel anxious because you feel like- It's just like - yeah, 'cause we don't - I feel like I don't know really know what's going on. [Laughs] (V1 midcourse)

Again V1 reiterates that students felt pressure to get the correct answer: "it's like, man, am I supposed to know this?" She then talks about how this pressure causes anxiety, especially when she feels that she is unsure whether this is something she is "supposed to have learned" or not. She tries to compare the instructor's expectations when using PI with what she feels is expected in a traditionally-taught course, where "new concepts come up in class." Therefore, a great deal of V1's negative experience appears to be related to her desire to meet the instructor's expectations. Departing from her previous experiences in lecture, she feels that the expectations are unclear, creating additional anxiety when she is unable to answer a ConcepTest correctly.

While both male and female students spoke about feeling pressure to get the right answer, previous research on how testing environments can impact stereotyped groups, such as females in science, suggest that the impact of ConcepTests may be differential for males and females. In the next section, we discuss how this literature may shed some light on the gender differences in student responses to PI with respect to these findings.

4.5 Discussion

Our findings from the student interviews align with the results found by Laws et al. (1999) at Dickinson College. Just as several of the junior and senior female students in Workshop Physics responded negatively to the social aspects of the collaborations in small group work, e.g., personality differences, concerns about their peers' perceptions of them, we find that our female interviewees also had much to say about the social issues that arose during PI. These issues, such as needing to navigate social relationships with a teaching relationship, were not only emphasized more by the female students than the male students we interviewed, but also appeared to impact their participation in the peer discussions. Although Jones et al. (2000) examined elementary student interactions in small group work, they also found that female students' emphasis on preserving social relationships interfered with their participation during discussions.

As Redish (2003) states, most physics instructors have goals for their students beyond just assimilation of content, but also goals related to what it means to “make sense” of this material. Instructors, often tacitly, expect students to make connections between concepts, draw on their personal experiences, recognize and evaluate assumptions, and relate and apply different representations. The results presented here, in conjunction with previous research, suggest that students, specifically female students, also may have a “hidden” agenda in these introductory physics courses—a need to maintain their social relationships. When asking students to interact with each other instructors may not pay attention or recognize that this interaction not only involves discussing content, but navigating friendships and other relationships with their peers.

The students we interviewed did not mention the status issues that often impacted who participated in group work in previous studies, e.g., domineering male partners or feeling ignored because of minority status in the course. It could be because students formed their own groups in this course and there were approximately equal numbers of both genders that these issues of minority status may have been less relevant. However, for some students, the peer discussions often took the form of one student teaching the other student the correct answer, which automatically placed the teacher-student as higher status over their peers. One female interviewee in particular observed this status imbalance and noted that she did not participate as much in the discussions because she did not want to seem like she was “teaching” or giving herself higher status among her peers.

Related to this issue of status in discussions, both male and female students reported feeling pressure to answer the ConcepTests correctly, especially in comparison to their peers. Although students are not given credit for correct answers, they perceive a testing environment when they answer the clicker questions individually. As discussed in the background section, a testing environment can have particularly negative consequences for stereotyped groups, such as female students in a physics course. Even though both male and female students report feeling stress to obtain correct answers, female students may feel additional stress to answer correctly so they do not confirm a negative stereotype about females and physics. While we did not find evidence that female students are responding to stereotype threat in this PI environment, we would not likely obtain this type of evidence during student interviews as the effects of stereotype threat may not always be not conscious (Croizet & Claire, 1998; Oswald & Harvey,

2000). It is possible that the emphasis that students place on answering these ConcepTests correctly may be related to the more negative responses that females have about PI.

We must be extremely cautious not to generalize these results beyond this course under study and the student population that participated in the interviews. We only interviewed a few students from the course, most of whom obtained the very high grades in the course; therefore these results may not be representative of all students in the course. Yet, the quantitative survey results also indicate that there are gender differences in how students respond to this PI environment. In addition, we also are limited by the fact that the students enrolled in this course attend a highly-selective university. Therefore, these students are not representative of all introductory physics students nationwide and furthermore, the institutional context in which this course and study are situated is distinctive. These unique attributes of our study provide reason for caution in interpreting these results more broadly. For example, the testing atmosphere that the students report is likely influenced by the competitive nature and norms of the institution that extend beyond this course. However, as discussed in the background section of this chapter, similar findings have been found at other institutions, at the post-secondary, secondary, and elementary school levels, suggesting that elements of the themes we describe here can be found in a variety of contexts. These results all point to the need for further research on how all students, but particularly under-represented minority students, experience interactive engagement classrooms.

While many of the students in our study and others bring up negative aspects to the peer discussions, we do not think that the solution is to remove the interactive aspects

of the pedagogy. Working in small groups is a skill needed not only for professional scientists, but for all professionals. Not only do scientists need to present and defend their work in journals, conferences, and other public media, but the scientific concepts and theories themselves are socially constructed, refined, and accepted within the scientific community. Therefore, to learn about science, students must learn how to engage in scientific discourse and argumentation. PI and other interactive teaching techniques give students the space and opportunity to practice discussing and arguing their scientific ideas. However, just as the learning of science content needs to be scaffolded for the students, so does the learning of the practices of scientific discourse and argumentation. In a PI classroom, students first discuss with each other, but there are opportunities for the instructor to model and scaffold scientific arguments for the students. For example, the instructor can lead a whole-class discussion after the peer discussions, asking students to put forth arguments for different answers (not just the answers they believe to be correct) and encouraging other students to build on or respond to these ideas (Turpen & Finkelstein, 2009). By modeling and guiding students in scientific discourse and argumentation, the instructor will support student learning about not just the content, but the practice of science and help align students' expectations about the peer discussions. Many of the interview comments in our study were about navigating relationships and the social atmosphere; establishing norms for discourse may help prevent these social issues from arising during peer discussion.

The actions of the instructor have already been shown to impact how students behave and perceive PI. For example, Lucas (2009) found that when given no instructions on how to discuss with their peers, only a few dominant students participated

in the peer discussions. However, when given explicit instructions to first talk to their neighbor using paper and pencil, then share their ideas with their group; more students participated in the discussions. At University of Colorado, Turpen and Finkelstein have identified different features of PI implementation that result in different classroom norms and student perceptions of these norms (Turpen & Finkelstein, submitted; Turpen & Finkelstein, 2009). Their results may be particularly informative on the classroom norm of *answer-making* versus *sense-making* (diSessa, et al., 2002). When *answer-making* students are engaged in the activity of finding the answer and explanation that they think the teacher wants to hear. In contrast, *sense-making* involves seeking out an explanation that makes sense to the student, by making connections to their previous ideas and experiences. Turpen and Finkelstein (2009) claim that the actions of the instructor send a message to the students about how they should frame the clicker questions and discussion. For example, if after peer discussion the instructor simply explains the answer to the ConcepTest, or listens to only one students' correct explanation, the norm of *answer-making* is supported. In contrast, listening to multiple students' explanations, some correct and some incorrect, and helping students build on these ideas promotes the norm of *sense-making*. In light of the emphasis that our interviewees placed on obtaining the correct answers to ConcepTests, implementation changes such as leading a whole-class discussion of multiple students' ideas may alleviate some of the testing atmosphere to which students negatively responded.

4.6 Conclusion

In the previous chapter, we discovered that female students rate the teaching methods, specifically PI, lower than male students in an introductory physics course. Through mid-course and post-course interviews, we find that the experiences of male and female students in this PI environment may be different, due to gender differences in attention to the social issues that arise during peer discussion and responses to testing environments. These results emphasize the need to better understand *how* students experience these interactive reforms, particularly in large lecture classrooms. Furthermore, we hypothesize that implementation features which help scaffold students in the scientific practices in which they engage in PI (e.g., sense-making and scientific argumentation) may address some of the gender differences in student response to interactive teaching. We discuss these features in further detail in chapter 6, when we assimilate the implications for instruction for each of the three studies described here.

Chapter 5

Gender, mental rotations, and introductory physics

5.1 Introduction

In chapter 2, we showed how the gender gap in performance in introductory physics courses is mediated by the use of interactive pedagogies. However a gap still remains, especially in the course for bioscience majors. In this chapter, we examine an often cited reason for this gap in achievement and the observed disparity in participation in upper levels of physics: differences in assessments of cognitive abilities explain differences in achievement and participation. More specifically, differences in cognitive assessments of *spatial ability* are cited as possible explanations for the gender gap.

A recent review paper by Taasobshirazi and Carr (2008) looked at gender differences in physics through the lens of the expert-novice framework. They use this framework to hypothesize why different factors, such as cognitive differences or social influences, relate to gender differences in achievement and participation. In their analysis they suggest that cognitive differences in spatial ability may be a factor in how students use representations. If male students have higher spatial ability, and this results in more expert-like use of representations, this advantage may account for gender differences in science and engineering. To address this hypothesis, in addition to examining the quantitative relationship between performance on a spatial ability test and performance in an introductory physics course, we also examine student problem-solving in detail. We

investigate how students with different spatial ability scores approach highly-spatial introductory physics problems and how they use different tools, including representations, to help them solve these problems.

In this project, we aim to build on the previous literature concerning how sex differences in cognitive abilities relate to achievement in the sciences, but from a socio-cultural perspective. As such, our research questions are framed not only to add to the research by correlating performance on mental rotation tests with performance in a science course, but also to consider the internal (cognitive) and external tools students use to solve problems in a physics environment. In this chapter, we address the following questions:

- (1) Do students who perform better on an assessment of mental rotation abilities perform better in an introductory physics course? Furthermore, how do gender differences in performance on the mental rotations test map on to gender differences in physics achievement?
- (2) What tools do students use to solve highly-spatial physics problems?
- (3) How do students use mental-pictorial representations and transformations as specific tools in physics problem-solving?

5.2 Background

5.2.1 What cognitive sex differences have been documented?

Many researchers have looked at sex differences in cognitive abilities—but recently more emphasis has been placed on the similarities between males and females in

reasoning skills important for science and mathematics (Hyde, 2005; Spelke, 2005). In a review of meta-analyses of psychological sex differences, Hyde (2005) found that males and females had very few or no differences on more than three-quarters of the cognitive variables studied, including those typically associated with success in math and science. Surprisingly, differences between the sexes on mathematical ability and problem-solving were small ($d = 0.08$ – 0.16) (Hyde, 2007). On tests of spatial perception and visualization, the effect sizes were small to moderate ($d = 0.13$ – 0.44) and the magnitude of the sex differences differed between studies (Voyer, Voyer, & Bryden, 1995). However, there were consistent sex differences on tasks involving mental rotation, on which males outscored females with effect sizes ranging from 0.56 – 0.73 (Hyde, 2005). These effect sizes were moderate to large and the differences between males and females on mental rotation tasks are consistent throughout the literature (see the following meta-analyses for more details: Halpern, 2000; Linn & Petersen, 1985; Voyer, et al., 1995).

5.2.2 What are mental rotations and how are they assessed?

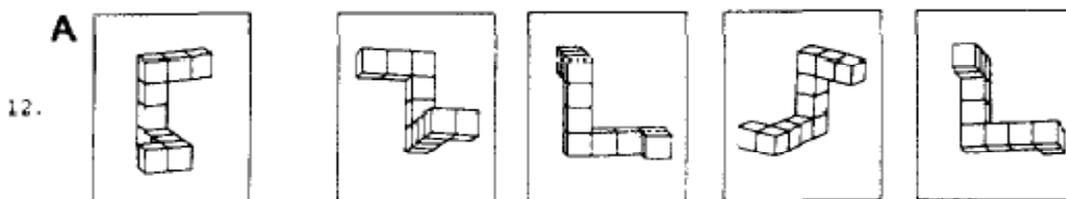
Mental rotation is the process of forming a mental image of two- and three-dimensional figures and rotating the image in a variety of orientations. Shepard and Metzler (1971) first proposed this cognitive process after asking individuals to compare two two-dimensional images of three-dimensional objects at different orientations. The time that participants took to determine whether these images were of the same object was found to be a monotonic linear function of their angular difference, which the authors took as evidence that individuals rotated mental images to carry out their comparisons. Follow-up studies with different objects and test instructions confirmed the

linear relationship between angular disparity and time needed to compare two objects (for a review, see Shepard & Cooper, 1982); however different experimental procedures and stimuli appear to affect the value of the slope (e.g., Ark, 2005; Cooper, 1975).

Mental rotation assessments

Using the same block objects as in the original Shepard-Metzler paper, the Vandenberg and Kuse Mental Rotations Test (MRT) (M. Peters, et al., 1995; Vandenberg & Kuse, 1978) requires subjects observe a target figure, and determine which two of the four stimulus figures are rotations of the target. Figure 5.1 contains an example from one version of the re-drawn MRT (M. Peters, et al., 1995). This is a timed assessment split into two sections; students are given three to four minutes to answer twelve items.

Figure 5.1. Sample item from the redrawn Vandenberg-Kuse Mental rotation test (Vandenberg and Kuse, 1978 and Peters, et al., 1995). Students are asked to determine which of the four figures on the right are rotations of the target figure on the left. In this problem the second and third figure are rotations of the target figure.



While there are many different assessments of mental rotations, the MRT obtains the largest and most consistent sex differences in both overall performance and accuracy (M. Peters, et al., 1995). Another test, the Purdue Visualization of Rotations Test (ROT) (Bodner & Guay, 1997) was developed and used to investigate the relationship between mental rotation ability and performance on chemistry assessments. A sample question

from this test is displayed in Figure 5.2. In contrast to the MRT, these problems are multi-step: subjects must first observe how a target is rotated in two figures and then map that rotation using a different shaped stimulus. The authors also note that defining characteristics can be hidden from view, preventing subjects from just mapping key features and not performing whole figure rotations. Additionally, the rotations use differing axes that are not limited to the plane of the page or the axis normal to the page (Bodner & Guay, 1997).

Figure 5.2. Sample item from the Purdue Visualization of Rotations Test (Bodner & Guay, 1997). Students are asked to replicate the first rotation using a different figure. For this question, choice B is rotated from the target figure to mimic the first rotation.

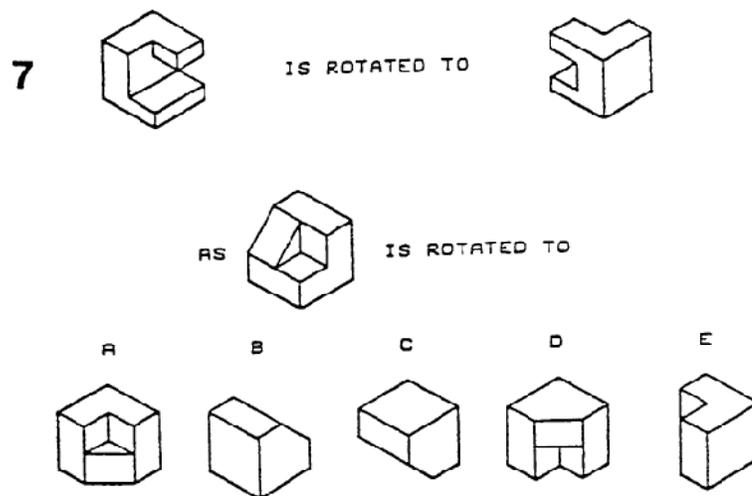


FIGURE 1. ITEM 7 FROM THE 20-ITEM VERSION OF THE PURDUE VISUALIZATION OF ROTATIONS (ROT) TEST. COPYRIGHT, PURDUE RESEARCH FOUNDATION.

Explanations for gender differences and impact of strategy and training

Both biological and social factors have been cited as causes of the sex differences on mental-rotation tasks (e.g., McGee, 1979). Researchers who claim that biological factors account for sex differences point to sex hormones (e.g., Hausmann, Slabbekoorn,

Van Goozen, Cohen-Kettenis, & Güntürkün, 2000), genetics—specifically x-linkage (Bock & Kolakows, 1973)—and different brain lateral organization (e.g., Rescher & Rappelsberger, 1999). From the environmental perspective, researchers have found that males are more exposed than females to spatial activities and training, such as outdoor sports and video games, and that exposure to these spatial activities is related to performance on mental rotation assessments (for meta-analysis and review, see Baenninger & Newcombe, 1989, 1995). Interestingly, the sex differences in performance on mental rotation tasks are observed to be culture-independent, with similar effect sizes across cultures (Mann, Sasanuma, Sakuma, & Masaki, 1990; M. Peters, Lehmann, Takahira, Takeuchi, & Jordan, 2006).

While the MRT and similar tests claim to measure ability to perform mental rotations, results on strategy use suggest that these tests could also be measuring other aspects of student thinking. Peters, et al. (1995) asked participants about the different strategies that they used to answer questions on the MRT and found small, but statistically significant sex differences in reports of the strategies used. A greater percentage of females reported using verbal or analytical strategies, i.e., thinking through the comparison verbally or using their hands or physical objects to help them judge whether two objects at different orientations were the same. For both genders, individuals who used primarily *nonverbal* strategies obtained higher average scores on the MRT than those who used verbal strategies. Several follow-up studies provide further evidence of the use of multiple strategies on the MRT and other mental-rotation assessments, but the results on sex differences in strategy use are mixed (Burin, Delgado, & Prieto, 2000; Geiser, Lehmann, & Eid, 2006; Li & O'Boyle, 2008; Schultz, 1991). Studies using fMRI

to examine brain activation in males and females performing these mental rotation tasks also suggest the use of different strategies based on cerebral activation patterns (e.g., Ark, 2005; Hugdahl, Thomsen, & Ersland, 2006; Jordan, Wüstenberg, Heinze, Peters, & Jäncke, 2002; Weiss, et al., 2003).

Glück, Machat, Jirasko, and Rollett (2001) found that training can change strategy use and improve performance on the MRT. Further research shows that training on spatial tasks can improve individual scores (e.g., Leone, Taine, & Droulez, 1993), but the results on the effects of training on sex differences are more mixed. For example, a meta-analysis by Baenninger and Newcombe (1989) found that while training can improve the scores of both males and females, there remains a persistent gap between the sexes. However, more recent results suggest that video-game training can preferentially improve female performance on mental rotation tasks (e.g., Feng, Spence, & Pratt, 2007). Furthermore, some studies show that practice rotating the same figure, such as the Shepard-Meltzer mental rotations cubes, can not only improve scores on assessments using the same figures and reduce the sex differences in performance, but that this improvement can transfer to other stimuli (Wright, Thompson, Ganis, Newcombe, & Kosslyn, 2008) as long as the task involves similar processing demands (Wallace & Hofelich, 1992).

In summary, sex differences on mental rotation tasks are robust; however research on the causes of and solutions for these differences is not conclusive. Interestingly, Kozhevnikov and Thornton (2006) found that physics instruction, especially instruction involving the use of microcomputer-based laboratories, improved participants' performance on spatial tests, including the MRT, although the effect on sex differences

was not reported. In the next section, we further examine the relationship between mental rotation skill and the participation in scientific disciplines.

5.2.3 What is the relationship between mental rotation and science learning?

In his book outlining his theory of multiple intelligences, Howard Gardner suggests that “it is skill in spatial ability which determines how far one will progress in the sciences” (Gardner, 1993). Indeed, students who demonstrate higher spatial ability prefer math and science classes (Lord & Nicely, 1997) and are more likely to pursue careers in science, math, and engineering (Shea, Lubinski, & Benbow, 2001). In physics in particular, mental visual imagery is thought to be integral to the development of Einstein’s theories of relativity, Faraday’s contributions to electromagnetism, and Bohr’s model of the atom (Botzer & Reiner, 2005; A. I. Miller, 1984, 2000).

General spatial ability and physics performance

Pallrand and Seeber (1984) studied the relationship between visuo-spatial abilities and achievement in students in introductory physics courses. They found that high gains on a test of visuo-spatial skills were related to test items involving graphs and laboratory work. Furthermore, students who withdrew from their course tended to have lower spatial abilities, but similar math skills, to students who remained in the course. In the upper division, Browne (2002) examined the relationship between performance on a test of spatial visualization and grades in advanced physics courses and found no statistically significant correlation. Perhaps this finding is not surprising, since final grades often take many aspects of a course into account, such as effort or participation, and may not be the best measure of understanding or ability. Interview data from the

same study did suggest that visualizations play an important role in physics problem-solving, and that junior physics majors already have high spatial ability in three dimensions (Browne, 2002).

Several researchers have suggested that students with high spatial ability may be able to better process different components or dimensions in physics problems. Isaak and Just (1995) found that students with high spatial ability, as measured by the Space Relation Test, a subtest of the Differential Aptitude Tests (Bennett, Seashore, & Wesman, 1972 from Isaak and Just, 1995), were able to more accurately describe the trajectory of a point on a rolling wheel. To describe this rolling motion, the authors contend that students must simultaneously process both the translational and rotational motion of the wheel. They suggest that both describing rolling motion and tests of spatial ability depend on students' abilities to store and process visuo-spatial information. Similar results were found in introductory physics problems in two dimensions (Kozhevnikov, Motes, & Hegarty, 2007). When solving two-dimensional kinematics problems, the authors found that high-spatial students were more likely to account for and integrate both the horizontal and vertical components of motion.

Mental rotation and physics performance

While the studies described above examined relationships between physics problem-solving and spatial skills in general, they did not specifically address skill in mental rotation, which is, as described in the previous section, the spatial sub-ability with consistent sex differences on assessments. Kozhevnikov, Hegarty, and Mayer (2002) correlated performance on a kinematics questionnaire with performance on various assessments of spatial ability. While there were statistically significant relationships

between kinematics scores and performance on tests assessing spatial visualization and spatial orientation, the speeded rotation tests, similar to the MRT, were not related to any type of kinematics problems.

Chabay and Sherwood (2006) suggest that mental rotation ability is important for students learning electricity and magnetism, which requires the use of cross-products and reasoning in three-dimensions. Exploring the relation between mental-rotation ability and understanding of electromagnetism, Sağlam and Millar (2006) only found a weak correlation in upper secondary students in Turkey. However, students did not have much variability in the scores on their mental rotations test (ROT), with many students achieving the maximum score. Furthermore, in interviews, students' problems appeared to be very basic—such as on the ontology of the magnetic field—and it is unclear how mental rotations would be a factor in their performance. This study suggests only a weak correlation between electromagnetism understanding and performance on mental-rotations tasks; however it is unclear what the relationship would be for students with more advanced content knowledge and with a more appropriate assessment of mental rotations.

Mental rotation and chemistry performance

Several studies in chemistry also found a relationship between mental rotation ability and performance in introductory courses (Bodner & McMillen, 1986; Carter, Larussa, & Bodner, 1987). These correlations were observed in both highly spatial tasks (e.g., identifying crystal structures) and seemingly non-spatial tasks (e.g., stoichiometry problems) (Bodner & McMillen, 1986). However, differences in correlation were found to be dependent on the complexity of the tasks; success on tasks requiring rote

memorization or simple algorithms was not related to mental-rotation test scores. Pribyl and Bodner (1987) elaborated on the relationship between student spatial abilities and chemistry problem-solving, finding that students with high spatial ability tended to draw preliminary drawings and achieved more success on more complex (multi-step, open-ended) and rotation-intensive problems in an organic chemistry course. One explanation is that the test used in these studies (ROT, described in the previous section) uses multi-step questions, making it possible that it is not mental rotation processes that are correlated with performance, but the cognitive processes involved in multi-step problems. Furthermore, in a review on these and other studies, Wu and Shah (2004) suggest that general reasoning skills, which are correlated to spatial abilities, could be the more important factor in predicting academic performance. Stieff (2004) provides evidence for this hypothesis, finding that the correlation between chemistry performance and visuo-spatial ability is mitigated when controlling for SAT math and verbal scores.

Stieff (2004, 2007) further challenges the assumption that visuo-spatial strategies are critical for high achievement in chemistry by analyzing students' problem-solving on highly spatial chemistry problems. He finds that students can use analytical strategies to solve chemistry problems, such as comparing molecular structures, which may obviate the need for visualization. In addition, experts in chemistry problem-solving were found to use a range of visualization and analytical strategies on organic chemistry problems, but, on average, preferred to use basic algorithms and heuristics (Stieff & Raje, 2008). These results suggest that chemistry problems that appear to require imagistic reasoning can be solved without using spatial visualization techniques, thus minimizing the role of

individual differences in spatial cognition for successful performance in chemistry courses.

In summary, while there is extensive research examining the relationship between students' spatial abilities and science learning, these results seem to vary based on the specific spatial assessments employed and science content studied. Furthermore, few of these studies focused on mental-rotation assessments in particular. Therefore, though spatial abilities have been shown to be instrumental in physics discoveries and related to physics problem-solving, the relationship between mental-rotation ability and physics is less clear. Although males perform consistently better on tests of mental rotation, it is unclear whether this difference is related to performance in physics, where sex differences occur starting in late high school (Hazari, et al., 2008). The ability to perform mental rotations is hypothesized to be critical for learning in electricity and magnetism, however the only study relating performance on a mental rotations test and performance on an electricity and magnetism test found a small relationship and these results were not conclusive. There is a need for additional research to better evaluate the claim that sex differences on tests of mental rotations explain sex differences in physics. In this chapter, we address this need by looking at introductory physics courses at an elite four-year university.

5.3 Conceptual Framework

Spelke (2005) suggests that gender differences on tasks involving spatial skills are often based on how males and females use different strategies to solve problems. She cites research that shows that, on problems that can be solved using verbal computation or

spatial visualization methods, males are more likely to use spatial imagery and females are more likely to use more verbal computation methods. Furthermore, on mathematical problems that are more easily solved using spatial imagery, males tend to perform better, but when all students are encouraged to use the spatial strategy, the gender gap is narrowed. As mentioned in the above section, students even use different strategies on the MRT (M. Peters, et al., 1995) and alternative methods of priming on mental rotation tasks have been found to affect strategy use (Kanamori & Yagi, 2005). These results on the different strategies students use, and the conditions that encourage different strategies, suggest that experience, context, and culture all play important roles in problem-solving. As Shore (1996) states regarding mental rotations and transformations, “Though the human sensorium seems to be genetically prepared for such visual acrobatics, actual feats of perception must be brought to life through an individual’s concrete interactions with the world.” (p. 4)

Many researchers have documented the importance of culture in problem-solving. For example, Nunes, Schliemann, and Carraher (1993) found that children of street vendors in Brazil used different strategies to do similar arithmetic problems depending on whether they were in the marketplace or in a formal school setting. When asked how much 5 coconuts would cost in the marketplace, the children would manipulate the quantities in their head, using a series of smaller intermediate multiplications and additions. On the same multiplication question on a formal test, the children would attempt to use the school-prescribed written routines involving manipulation of symbols, with much less success. Students performed arithmetic differently depending on the setting, highlighting the importance of context and culture in assessments of problem-

solving. Similarly, Cole (1996) found that culture can have a much more fundamental role in cognitive processes. Western teachers visiting Liberia believed that Africans had many debilitating intellectual difficulties, especially with mathematics. In traditional Western classroom settings, these assumptions were confirmed, but when the tasks were adapted to fit within their indigenous cultural practices, dramatically different results were found. For example, although Liberians were thought to be unable to work with measurements based on assessments using traditional classroom tools, when using indigenous standards of measurement for rice, specifically a tin can called a *kopi*, the adults in Liberia outperformed American adults at estimating large and small amounts of rice. Cole (1996) claims that these results show the importance of cultural tools in relation to cognitive skills. Both examples show how culture can provide tools that help people employ new actions, but may also limit their actions when solving a problem.

Wertsch (1998) gives another multiplication example to demonstrate how cultural tools facilitate mathematical problem-solving. In multiplying 343 by 822 (without a calculator), most people would use a vertical array in their calculations:

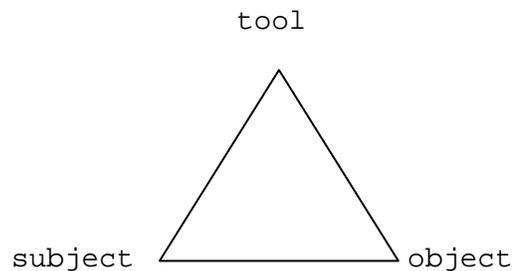
$$\begin{array}{r}
 343 \\
 822 \\
 \hline
 686 \\
 686 \\
 2744 \\
 \hline
 281946
 \end{array}$$

Wertsch posits that it is not just the person who solves this multiplication problem, but the person plus the written algorithm. The vertical array is a tool that mediates the solving of this problem. These ideas about how we use tools to help negotiate our actions originate in work by Vygotsky, who claimed that to understand

students' mental processes, we must understand the tools and signs that mediate them (Cole & Wertsch, 1996). Figure 5.3 depicts this idea of mediation, showing how individuals relate to their environment both directly and indirectly through the use of tools (Cole & Engestrom, 1993).

Vygotsky (Vygotsky, 1981 from Wertsch, 1998) noted that many internal and external devices could be tools: “language; various systems for counting; mnemonic techniques; algebraic symbol systems; works of art; writing; schemes, diagrams, maps, and mechanical drawings; all sorts of conventional signs; and so on.” In physics, Otero (2004) notes the use of the napkin as a tool for communication and a tool for thought in stereotypical back-of-the-napkin calculations. Learning what tools students use and how they use them can provide insight on how students make meaning of physics concepts and problems.

Figure 5.3. Depiction of the direct interaction between a subject and an object, as well as the indirect interaction mediated by a tool.



The idea that we should look at what tools students use to help them learn physics content or solve physics problems is not new in physics education research. For example, use of external representations in problem-solving has been well-documented and

researched in physics education (e.g., Chi, Glaser, & Rees, 1982; Kohl, 2007; Meltzer, 2005; D. R. Rosengrant, 2007; Van Heuvelen, 1991). Individuals can use pictorial, diagrammatic, graphical, and mathematical representations as tools that mediate their interactions when solving a problem or learning a new concept (Kohl, 2007). Van Heuvelen (2001) suggests that representations can be referents that can help provide meaning to abstract words and symbols as students are learning physics concepts. For example, Feynman diagrams are often constructed to help calculate the scattering cross-section in quantum mechanical processes. This diagrammatic representation mediates the solution to a problem in relativistic quantum mechanics by acting as a concrete referent for more complicated mathematical representations (Van Heuvelen, 2001).

In this project, we view mental rotations and/or transformations as an internal tool that people can use on the MRT as well as on physics problems. For example, if asked whether a frog has a tail, most people report generating a mental image of a frog and “mentally rotate” this image to “see” the frog’s backside (Kosslyn, 1994, p. 327). In this situation, we consider the mental-pictorial representation and transformation as tools that people use to think about the question.

By viewing mental rotations as a tool, we depart from the traditional view of mental rotations as an ability that is measured by the MRT. Just as studies in mathematics education and psychology found that people who were not able to display an ability or answer a question in one context but could demonstrate the same ability in another, we cannot draw conclusions on whether students have an ability to do mental rotations based on their performance on one assessment. As seen in the literature, students may choose *not* to use three-dimensional mental transformations on MRT questions or use a

combination of strategies. Students may have more practice or skill in using this tool, which can be seen in the mental rotations literature: the MRT is highly susceptible to practice effects (e.g., Wright, et al., 2008) and students may have more or less experience using mental rotations in other contexts, e.g., video games (Feng, et al., 2007). Therefore, instead of focusing on whether or not students ‘possess’ mental transformations as an ability, in our interview analysis we chose to focus on whether they use mental transformations in physics contexts and how successfully they use this tool in their problem-solving. Furthermore, as some students may not be as successful at using mental rotations tools or may choose not to use this tool, we examine what other tools students can use and how they use them on highly spatial physics problems.

5.4 Methods

To answer our research questions, we used a mixed-methodology research design. We collected survey data to examine the relationship between student performance on a test of mental rotations and performance in an introductory physics course data. For a more in-depth look at how students used mental transformations in physics problem-solving, we conducted a series of problem-solving interviews, asking students to talk aloud as they solved highly spatial physics problems in electricity and magnetism. Both the survey and interview data are used to answer the first research question, while interview data are used primarily to answer the last two questions.

We conducted this study in two introductory physics courses at Harvard University, collecting data from:

- Physical Science, a course for biological science majors and
- Honors Physics, a course for physics and engineering majors.

We chose these introductory physics courses because they cover topics in electricity and magnetism; these topics are hypothesized to relate to mental rotation performance (Chabay & Sherwood, 2006). Physical Science is one course in an integrated physical and chemical sciences sequence designed for students who are planning to major in the life sciences. Emphasizing relevance to biological systems, the course covers topics relating to light and sound waves, electricity and magnetism, and the transmission and processing of information in electronic and biological systems. Honors Physics is the second-semester course in a sequence for physics and engineering majors, covering topics in electricity and magnetism using the textbook by Purcell (1985). Students in Honors Physics were expected to have taken or be concurrently enrolled in a vector calculus course; operators such div , grad , and curl were used in the course. Considering the nonstandard nature of both courses (Physical Science and Honors Physics), as well as the exceptional background and preparation of students at Harvard University, we cannot consider these data as representative of introductory physics students nationwide. However, our analysis provides insight on how performance on the mental rotations relates to performance in two very different courses and shows how students with different preparations use mental rotations in physics problem-solving.

5.4.1 Quantitative data

Course surveys and achievement data

To examine the relationship between performance on mental-rotation assessments and performance in introductory physics courses, we asked students to complete two assessments and collected their course achievement data, as listed below:

- Vandenberg-Kuse Test for Mental Rotations
- Conceptual Survey on Electricity and Magnetism (both overall and magnetism-specific)
- Final exam scores (both overall and magnetism-specific)
- Final grades

We administered the Vandenberg-Kuse Test for Mental Rotations (MRT) on paper within the first two weeks of class. In Physical Science, we asked students to complete this assessment during their first laboratory. In Honors Physics, students took the MRT at the end of a lecture in the second week of class. The MRT consists of two sections, both with 12 questions, and we gave three and a half minutes to complete each section. Students reported additional demographic information (gender, race, year in school) on an attached sheet after completing the assessment. The MRT has an estimated internal-consistency reliability of 0.83 and research gives evidence of its construct validity (Vandenberg & Kuse, 1978).

We administered the Conceptual Survey on Electricity and Magnetism (CSEM, Maloney, et al., 2001) online to students as a pretest and posttest. During the first two

weeks of class and again two weeks before the final exam, students logged on to the test website with their name, ID, and email address. Students had one hour to complete the 32-item test, without the use of outside materials, and received credit for participation but not for correctness. The CSEM has been used to assess conceptual understanding of electricity and magnetism in introductory courses across the country. The overall test has an estimated internal consistency reliability of about 0.75 and individual questions were validated by 42 two-year college professors (Maloney, et al., 2001). To minimize practice effects, students did not receive any feedback on their responses to the pre-course CSEM and they did not receive copies of the test after either administration. We analyzed data using students' overall score, as well as their score on questions on magnetic field and force. We obtained similar results in both analyses; therefore we present here only the results using students' total scores.

We obtained course-achievement data from students in both Physical Science and Honors Physics, including final exam scores and final grades out of 100. Because Physical Science covered many diverse topics not directly related to electricity and magnetism, we also looked at student performance by question on the final exam, isolating a question involving magnetic induction. We found similar results whether we used the entire exam or this individual question as the outcome variable; therefore we present results using the complete final exam.

Description of our sample

The student sample in Physical Science ($N = 190$) was 70% female, while approximately 43% of the students in Honors Physics ($N = 58$) were female. During the MRT administration, students reported their major and year in school. In Physical

Science, 65% of students reported being science majors other than physics and engineering, while the remaining students were non-science majors. Most (70%) of the students were juniors, with a few sophomores and seniors also enrolled. More than half (57%) of the students in Honors Physics reported being physics or engineering majors, the remaining were other science (33%) or nonscience (10%) majors. Almost all of the students in Honors Physics were freshmen or sophomores.

Student participation on the MRT was optional, and although students did receive some participation credit for taking the CSEM pretest and posttest, not all completed these assessments. Therefore, we do not have complete datasets for all students enrolled in Physical Science and Honors Physics. In Table 5.1, we show the response rates for student participation on each of these surveys.

Table 5.1. Response rates on MRT, CSEM pretest and posttest, and all three combined for Physical Science (PS) and Honors Physics (HP).

	<i>N</i>	Response rate			
		MRT	CSEMpre	CSEMpost	All
PS	190	0.87	0.86	0.84	0.66
HP	58	0.66	0.81	0.69	0.55

To examine the differences between students with matched data and those with missing data, we present the comparison of Physical Science student scores, in Table 5.2. We report the mean score, standard deviation, and number of students for all students and for students with matched data and students with some missing data separately. To compare the average scores, we note the effect size (ES) of the difference between these two groups, and the associated *p*-value obtained when comparing the means of these two

groups. Students with complete datasets were about 31% male; similarly, 29% of students with missing data were male. Students who did not take either the CSEM pretest and/or posttest obtained similar average scores on the MRT as those who had complete datasets. Students with no missing data, however, obtained a higher average final exam score and final grade, with the latter difference statistically significant at the .05 level. Students did receive some course credit for completing the CSEM pretest and posttest, therefore some of the difference in final grades could be due to this impact.

Table 5.2. Scores for comparing Physical Science students with matched data to students who did not take either the MRT, CSEM pretest, or CSEM posttest.

	Matched			Missing			Comparison	
	<i>N</i>	Mean	SD	<i>N</i>	Mean	SD	ES	<i>p</i>
MRT	126	14.2	4.4	39	14.0	4.9	0.04	0.82
CSEM pre	126	9.6	4.5	37	9.6	5.1	0.00	0.98
CSEM post	126	18.0	5.7	34	17.4	7.3	0.09	0.64
Final exam	126	68.3	13.1	62	65.4	14.8	0.21	0.17
Final grade	126	82.0	7.1	62	79.6	8.6	0.30	0.05

Similarly, we show the comparison of Honors Physics students with matched and missing data in Table 5.3, again presenting the number of students, mean, and standard deviations. We also compare students with missing data to those with matched data by presenting the effect size and associated *p*-value, noting the significant differences between these two groups. Fifty percent of students with missing data were male, while almost two-thirds of students with complete data sets were male. Students with complete datasets obtained higher average scores on the MRT and all measures of course achievement, with statistically significant differences on the final exam and final grade.

Table 5.3. Scores for comparing Honors Physics students with matched data to students who did not take either the MRT, CSEM pretest or CSEM posttest.

	Matched			Missing			Comparison	
	<i>N</i>	Mean	SD	<i>N</i>	Mean	SD	ES	<i>p</i>
MRT	32	17.0	4.6	6	14.8	5.7	0.48	0.32
CSEM pre	32	18.0	6.1	15	15.1	8.0	0.48	0.17
CSEM post	32	26.1	4.3	8	24.1	6.7	0.47	0.30
Final exam	32	80.6	13.5	26	72.5	18.9	0.60	0.06
Final grade	32	87.6	7.0	26	81.5	10.7	0.87	0.01

In collecting MRT data, one of our research questions involves examining whether performance on the MRT explains differences in course achievement. In Table 5.4, we display the grades for male and female Physical Science students who took the MRT, as well as the grades for the students with missing MRT scores. The seven male students who did not take the MRT obtained the highest average scores of all the groups, while the seventeen female students without MRT scores obtained the lowest exam scores and grades. On the final exam, these female students scored lower than their male peers, on average ($p = 0.03$), and marginally lower than the female students who did take the MRT ($p = 0.08$). None of the differences on the final grade were statistically significant. The differences on the final exam indicate that the population of students who took the MRT was different, and not representative of the population of the entire course.

Table 5.4. Mean achievement scores (and standard deviations) for male and female Physical Science students, separated by participation on the MRT.

	Final exam			Final grade		
	male	female	<i>p</i> -value	male	female	<i>p</i> -value
MRT	69.6 (50)	66.9 (113)	0.22	82.4 (50)	80.8 (113)	0.20
no MRT	77.4 (7)	60.8 (17)	0.03	84.7 (7)	79.4 (17)	0.21
<i>p</i> -value	0.19	0.08		0.48	0.46	

Analytic methods

To answer our first research question, we estimate the correlation coefficients between MRT scores and scores on course exams and final grades. On course measures where there exist statistically significant sex differences, we will control for students' mental rotation test scores and conduct ordinary least-squares regression analyses, to ascertain whether differences on mental rotation tests predict differences in introductory course achievement. Finally, we will include students' CSEM pretest scores in our linear regression models, to control for students' incoming conceptual understanding.

Missing data pose multiple threats to our analysis. The data presented in Tables 5.2-4 suggest that ignoring students without complete datasets would produce misleading conclusions; the "complete-case" sample is not representative of the entire course (Little & Rubin, 1987). Furthermore, this "listwise-deletion" method of dealing with missing data has been widely discredited in the literature (e.g., Little & Rubin, 1987). In addition to list-wise deletion, other ad-hoc methods have been proposed to deal with missing data, each with severe statistical drawbacks (e.g., Schafer & Olsen, 1998). Mean substitution, which replaces the missing values with the average value across the sample on that variable, produces biased results, and artificially reduces the variance for the variable and diminishing its relationships to other variables. Furthermore, replacing the missing values using regression substitution, or any other method that imputes the values only once does not account for the error introduced by estimating the missing data. Therefore in our statistical analyses we use the method of multiple imputation (Rubin, 1987), to replace the missing values at random, multiple times, and then to account for the uncertainty by pooling the multiple slope estimates obtained in our regression analyses to examine the

relationships between gender, race, student expectations about learning, and their ratings of the teaching methods.

We use the method of multiple imputation to create several complete datasets using appropriate randomly imputed values for the missing data in each variable, based on the non-missing data we possess for individual students, so even if the data are not missing completely at random (i.e., the probability of missing data depends on one of the other variables such as final grade) we account for this bias (Wayman, 2003). Once we have these multiple datasets, we conduct identical regression analyses in each and pool the estimates. To obtain the overall estimate of a regression parameter, for instance, we average the individual estimates from each of the m imputed datasets. The variance (T) of the estimated regression parameter is estimated by the formula given by Rubin (1987):

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B$$

where \bar{U} is the average variance within each dataset and B is the variance between datasets. Because the total variance accounts for the differences between imputed datasets, the inferences we draw from these analyses—such as standard errors and p -values—account for the uncertainty from imputing the missing data. This method of analysis has been shown to work remarkably well with relatively few imputed datasets (Rubin, 1987) and has proven to be a “principled and statistically defensible” approach to dealing with missing data (Schafer & Olsen, 1998).

5.4.2 Qualitative data

Interview data

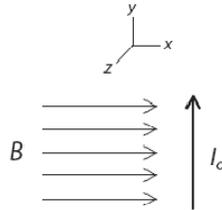
Through the surveys and course grades, we collected quantitative data on the relationship between students' mental rotations and physics performance, but to understand how students use mental rotations we must take a more in-depth look at students' problem-solving. We chose to look at how individual students solve problems, as their performance in an introductory course is most often determined by their performance on exams. We selected several students in Physical Science and Honors Physics to participate in think-aloud interviews toward the end of their course (e.g., either in the last month of class or in the reading period before the final exams, after they had covered the relevant material). We asked students to talk aloud as they solved five highly spatial physics problems and videotaped them as they thought through their solutions. While students talked through the solutions, the interviewer did not interrupt or ask students to describe or explain their thinking. By asking students to simply speak aloud what goes through their mind as they work, the interviewer is less intrusive on students' problem-solving processes (Ericsson & Simon, 1993) and the data may be more indicative of how students might approach these problems on their course exams.

To develop the interview problems, we modified questions from previous exams in addition to writing our own questions. We selected problems in electricity and magnetism that we thought would require reasoning in three-dimensions; all questions required some application of a cross-product. The third problem in the interview provided the richest data, as students solved this problem using myriad tools and with varied success. Therefore, we chose to base our analysis on students' solutions to this problem,

which we present in Figure 5.4. All problems and the interview protocol can be found in Appendix 4.

Figure 5.4. Problem 3 from student talk-aloud interviews.

An infinitely long wire is placed in a constant B-field, which is pointed in the x direction in all space.



- What is the direction of the force on the wire? What is the magnitude, in terms of force per unit length, using the parameters given?
- If the wire is rotated 30° clockwise about the x-axis, as seen from the right, what would the magnitude of the force on the wire be?
- Draw a graph of $|F|$ vs. ϕ , where ϕ is the angle of rotation of the wire about the x-axis.
- If the wire is rotated 30° clockwise about the z-axis, looking at the page, what would the magnitude of the force on the wire be?
- Draw a graph of $|F|$ vs. ϕ , where ϕ is the angle of rotation of the wire about the z-axis.

Interview participants

We purposefully selected students from the Physical Science and Honors Physics courses based on their pre-course CSEM and MRT scores, to obtain a sample of students with diverse physics backgrounds and performances on a mental rotations test. We invited 33 students who received high and low MRT and CSEM scores relative to their peers in the course; a total of 12 students accepted. Due to the low acceptance rate of Physical Science students with high MRT scores (e.g., in the upper 33%), we also selected a student with medium MRT score (Kyle). In total, we interviewed eight students from Physical Science and four students from Honors Physics. In Table 5.5, we summarize the scores and grades for the students who participated in the interviews.

Table 5.5. Survey scores and course grades for interview participants.

Name	course	gender	MRT	CSEM pre	CSEM post	final exam	grade
Gummy	PS	F	7	5	13	66	78
Green	PS	F	8	7	20	60	80
Jax	PS	F	9	5	14	74	84
Michelle	PS	F	9	6	24	90	92
Maverick	PS	M	10	30	26	97	97
Hawk	PS	M	12	15	23	82	90
Kyle	PS	M	15	22	24	82	90
Natalie	PS	F	20	22	28	98	100
Kelly	HP	F	11	13	21	78	85
Jack	HP	M	12	25	29	87	91
Matt	HP	M	23	10	21	59	79
Andrew	HP	M	24	25	31	97	98

Analysis of interview data

All interviews were videotaped and then transcribed, including students' verbal statements as well as physical gestures and drawings. To identify patterns in student problem-solving, we developed and applied a coding scheme inductively to identify and categorize the objects, concepts, operators, and other tools students used to help them understand or solve the problem. Using video footage, transcripts, and students' written work, we noted both the external tools, such as hand motions and diagrams, as well as the internal (cognitive) tools, such as the right-hand rule or trigonometry, that students indicated they were using when solving this problem. After developing these codes, two independent raters coded each of the transcripts, with 70% agreement between them before discussion and 97% agreement after discussion.

To address our third research question, we explicitly looked for, and coded student use of, mental-pictorial representations, using and building on the analytical

frameworks from Clement (1994) and Stieff and Raje (2008). We coded use of mental-pictorial representations either (1) when students stated explicitly that they were “visualizing” or “trying to picture” the problem or rotation, (2) when they used iconic or representational gestures (McNeill, 2000) and/or (3) when there was a gap in students’ reasoning that could not be explained through the use of any other tools the students were using at the time.

Regarding the first criterion, there were occasions in which students would make verbal claims about the use of mental imagery, but their gestures would indicate otherwise. For example, one student stated “*Okay I need to visualize exactly what’s happening here.*” [Kelly, quote 13], yet proceeded to use her hands for a three-dimensional representation of the problem (Figure 5.5). In these instances, we coded for student use of their hands, but not mental-pictorial representation, as the student was looking at and using external tools to ‘see’ the problem. While she could be using her hands to build a more complete mental picture, it is unclear from her words and gestures that she is using mental imagery as opposed to the physical representation. Therefore, we compared students’ verbal statements with their behavior on video to look for evidence that they were not primarily using mental-pictorial representations.¹⁶ If students were looking explicitly at and using physical representations, then mental-pictorial representation was not coded as a tool that students were using for that part of their solution.

¹⁶ Of course, at some level, we always have some internal picture of what is going on around us, so we are not arguing that students do not have these mental images. Rather, we are trying to piece together whether this is a tool they are explicitly relying on to help them solve the problem.

Figure 5.5. Use of hand motions that are not gestures, but physical representations of the problem. Although she claimed to be using mental transformations, it is not clear whether she is using her hands to help her transform her mental images or primarily relying on her physical representation to ‘see’ the rotation.



As opposed to hand motions that students deliberately use to build a three-dimensional representation of the problem, representational gestures are spontaneous hand movements that occur when speaking (e.g., Goldin-Meadow, 2005). Although gestures are mostly considered for their communicative value, research suggests that gestures play an important role in thinking about spatial information (Alibali, 2005). Furthermore, studies show that gestures help individuals maintain mental imagery and focus on spatial information (e.g., Wesp, Hesse, Keutmann, & Wheaton, 2001). Therefore, we noted students’ use of representational gestures and triangulated this evidence with their verbal statements to code for instances in their problem-solving that suggested use of mental-pictorial representations.

We also coded for use of mental-pictorial representations if students made leaps in their reasoning about the problem that could only be solved by manipulation of mental images, based on the other tools they were using at the time. These codes most often occurred when students were working on part (b) of problem 3 (Figure 5.4), when the wire needed to be rotated into the page. Students were often not using any other tools or only the pictorial representation and would come to the conclusion that the angle between the wire and magnetic field must still be perpendicular. For students to recognize that the angle is ninety degrees we postulate that they must be manipulating their mental image to extract that information; the use of the pictorial representation would not be enough. In the quote below, the student is not looking explicitly at any representation of the problem, yet determines the angle is ninety degrees.

Rotating around the X-axis. The B field's in the direction of the X-axis. So actually, contrary to what I was saying before, the B field and the current are actually staying perpendicular the whole time. No matter how you rotate around the X-axis. [Jack, quote 10]

Without the noticeable use of any other tools, the student appears to only be manipulating his mental image of the problem to solve for the magnitude of the force. Therefore, we coded for use of mental-pictorial representations in this instance and when there were other similar gaps in reasoning.

We note that this last criterion is actually using a lack of other tools as evidence of use of mental-pictorial representations. We developed this criterion to best explain how some students drew conclusions when their observable behaviors suggested that no other alternative strategy could be used. We could not determine any way for them to

determine that the angle between the wire and B-field was perpendicular without transforming their mental images to represent the rotation.

In addition to coding for the tools students used in each of the interviews, we also selected four student interviews for a more detailed analysis of how students used their tools on a highly-spatial physics problem. We selected two student interviews for case-study analysis based on the students' MRT scores and different use of mental-pictorial representations, providing additional insight into our first research question. The other two interviews were selected based on the extensive use of tools the students used. We compare and contrast these students' approaches to problem-solving, to provide additional insight into our second research question, by detailing how students coordinate their tools differently.

5.5 Survey Results

5.5.1 Gender differences in performance on MRT, CSEM, and course achievement

In Tables 5.6 and 5.7, we present summaries of the student results on the MRT, CSEM, final exam, and final grade for Physical Science (PS) and Honors Physics (HP) students respectively. We report the mean scores, standard deviation, and number of students for all students and for male and female students separately. To compare the average scores, we note the effect size (ES) of the difference between males and females, and provide the associated p -value for testing the hypothesis that the true value for the gap is different from zero, in the population.

Table 5.6. Summary statistics for MRT, CSEM pretest and posttest, and course achievement, by gender, for PS students.

	All students			Male students			Female students			Comparison	
	<i>N</i>	Mean	SD	<i>N</i>	Mean	SD	<i>N</i>	Mean	SD	ES	<i>p</i>
MRT	165	14.12	4.5	50	16.08	4.4	115	13.26	4.3	0.63	0.0002
CSEM pretest	163	9.63	4.6	48	10.80	5.4	114	9.16	4.2	0.36	0.04
CSEM posttest	160	17.84	6.0	47	18.83	5.9	112	17.51	6.1	0.22	0.21
Final exam	188	67.38	13.7	57	70.59	14.6	130	66.07	13.1	0.33	0.04
Final grade	188	81.22	7.7	57	82.71	8.0	130	80.60	7.5	0.27	0.08

Table 5.7. Summary statistics for MRT, CSEM pretest and posttest, and course achievement, by gender, for HP students.

	All students			Male students			Female students			Comparison	
	<i>N</i>	Mean	SD	<i>N</i>	Mean	SD	<i>N</i>	Mean	SD	ES	<i>p</i>
MRT	38	16.63	4.8	25	17.72	4.9	13	14.54	3.9	0.66	0.05
CSEM pretest	47	17.09	6.8	28	16.39	7.6	19	18.11	5.5	-0.25	0.40
CSEM posttest	40	25.73	4.9	23	25.65	5.7	17	25.82	3.6	-0.03	0.91
Final exam	58	76.95	16.5	33	74.18	17.1	25	80.60	15.3	-0.39	0.14
Final grade	58	84.87	9.3	33	83.11	8.9	25	87.19	9.4	-0.44	0.10

On average, male students in Physical Science obtained a statistically significantly higher score on the MRT, with an effect size of 0.63 ($p = 0.0002$). This difference is consistent with the literature on gender differences on this assessment (e.g., Linn & Petersen, 1985). Males also obtained higher average scores on the CSEM pretest ($p = 0.04$) and final exam ($p = 0.04$). The effect size indicates that on average, males scored about one-third of a standard deviation higher than female students. On the final exam, the gender difference is only marginally statistically significant ($p = 0.08$) and we found no differences on the CSEM posttest.

Males also obtained higher scores on the MRT, on average, than female students in Honors Physics ($p = 0.05$). The effect size (0.66) is again consistent with values found in other studies. Differences on the CSEM pretest, posttest, and final exam were not statistically significant. Females marginally outperformed male students, on average, on final grade ($p = 0.10$). The effect size indicates that females obtained average scores almost half of a standard deviation higher than their male peers.

5.5.2 Relationship between MRT and course achievement

In Table 5.8, we examine the relationship between performance on the MRT and course achievement for students in Physical Science and Honors Physics. We show the estimated correlation coefficients between MRT scores and scores on the CSEM pretest, posttest, final exam, and students' final grades. We also note the different sample sizes in each analysis; we used pairwise deletion to handle the occurrence of missing data.

In Physical Science, the estimated correlations indicate a small, but statistically significant relationship between performance on the MRT and performance in the course

($r = 0.18-0.22$, $p < .05$). Variability in MRT scores explains less than 5% of the variance in final grade. In Honors Physics, estimated correlation coefficients indicate a negative relationship between performance on the MRT and course achievement, although we cannot reject the null hypothesis that the estimated correlation coefficient is different from zero. Our sample is smaller in Honors Physics, thereby reducing our statistical power; however, the magnitudes of the estimated correlation coefficients are lower compared to those in Physical Science.

Table 5.8. Estimated pairwise correlations of MRT and CSEM, final exam and grade for students in Physical Science and Honors Physics.

	PS		HP	
	<i>N</i>	<i>r</i>	<i>N</i>	<i>r</i>
CSEMpre	140	0.15~	36	-0.12
CSEMpost	139	0.18*	32	0.08
final exam	163	0.21**	38	-0.11
final grade	163	0.22**	38	-0.13

~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

To investigate how MRT scores are related to course achievement and sex differences in Physical Science, we present multiple imputation estimates of fitted regression models for the final exam and final grade outcomes, in Tables 5.9 and 5.10, respectively. We first include gender (male) as a predictor in Model 1 and find that there are statistically significant differences between males and females on final exam ($p < .05$), and marginally statistically significant differences on final grades ($p < .10$). As expected, these results match the average gender differences presented in Table 5.6, with males outscoring females on average by 4.5 percentage points on the final exam and about 2 point on final grade. In Model 2, we include MRT scores, finding statistically

significant relationships between MRT and both final exam and final grade. However, when MRT is included, the differences between the sexes are no longer statistically significant.

Table 5.9. Multiple imputation estimates^a of parameters in regression models predicting scores on the final exam in Physical Science. ($N = 188$)

	Model 1	Model 2	Model 3
intercept	65.98	57.76	49.31
male	4.55*	2.75	
MRT		0.62*	0.47~
CSEM pre			1.21***

~ $p < 0.10$, * $p < 0.05$, *** $p < 0.001$

^aMultiple imputation conducted using proc mi in SAS 9.0.

Table 5.10. Multiple imputation estimates^a of parameters in regression models predicting scores on the final grade in Physical Science. ($N = 188$)

	Model 1	Model 2	Model 3
intercept	80.58	75.90	70.18
male	2.11~	1.12	
MRT		0.35*	0.26*
CSEM pre			0.77***

~ $p < 0.10$, * $p < 0.05$, *** $p < 0.001$

^aMultiple imputation conducted using proc mi in SAS 9.0.

In our final model for both final exam and final grade, we control for CSEM pretest scores, which is a statistically significant predictor in both models. Not surprisingly, the conceptual knowledge that students demonstrate before the course strongly predicts their achievement in the course. Once we control for this incoming knowledge, the predictive value of MRT scores is reduced for final exam; MRT only weakly predicts final exam scores in this model. Furthermore, although MRT remains

statistically significant in the final model for course grade, every one point difference in MRT scores is only associated with a *quarter* of a percentage point difference on final grade.

5.6 Results from student interviews

In the previous section, we found that there is a statistically significant, yet slight, positive relationship between students' MRT scores and course achievement in Physical Science, but no statistically significant relationship exists in Honors Physics. To better understand the relationship between mental rotations and physics problem-solving, we look to student problem-solving interviews on highly spatial physics problems. The results presented here are from analysis of students' thinking aloud as they solve problem 3 in their interview, which can be seen in Figure 5.5.

In this section, we first describe the tools we observed students using to solve this problem, including their mental-pictorial representations. In particular, parts B and C of this problem are highly spatial, involving rotation of the wire into the page and necessitating a three-dimensional picture to 'see' the rotation. Therefore, after describing the tools that students used on the entire problem, we examine which tools students used on part B specifically.

5.6.1 Describing the tools

Whether it was to understand the problem, plan a solution, execute the solution, or check their answer, we found that students used a variety of tools at each step in their

problem-solving process. We classify these tools into four categories and twelve distinct sub-categories, which are listed below:

Physics tools	Physics concepts/reasoning Right-hand rule Extreme values
Mathematics tools	Math calculation <i>including algebra, geometry, trigonometry</i>
Representation tools	Mathematical representation/equation Graphical representation Pictorial representation Physical representation Mental-pictorial representation
Other tools	Other parts of the problem/test Expectations Metacognition

The *physics tools* are categorized as tools that are taught in physics courses, although they are also tools that could be taught and used in different disciplines, including mathematics. The *mathematics tools* include all calculations using algebra, geometry, and trigonometry, which are all grouped together in a general subcategory of math calculations. Another mathematical tool is also documented: mathematical representations, which we note whenever students used or relied on a physics equation in their problem-solving. As students could perform math calculations separately from the equation, we keep these categories distinct. The manipulation or reorganization of an equation would incorporate both math calculation and mathematical representation tools. Inclusive of mathematical representations, the *representation tools* category also consists of use of physical, pictorial, graphical, and mental-pictorial representations of the problem that students used to guide or aid student thinking. The last category, *other tools*,

includes the remaining tools that we coded for in student interviews. In this section, we will give further details about each of the subcategories of tools, describing how we observed students use these tools and giving examples from interview transcripts and descriptions of the video.

Physics concepts

When students were talking through their problem-solving out loud, we observed them making statements that indicated that they were drawing on their background physics knowledge to help in their solutions. We found that several students recalled physics-specific concepts, including definitions, physical relationships, and models, to help them better make sense of the problem and their solutions. The quote below shows an example of one student who uses physics concepts immediately after reading the problem:

Okay, so first of all, we know that current is electrons moving or protons... movement of charge. It's in the direction of negative charge even if it's usually caused by positive charge. [Matt, quote 1]

To begin thinking about the problem, Matt turns to the physics he knows about the situation, “*current is... movement of charge.*” From this starting point, he then recalls (incorrectly) the convention of how to define the direction of the current. Matt appears to be trying to understand the problem by translating it into physics using these physics-specific concepts.

Natalie uses this same physics concept about electric current, but in coordination with the magnetic force equation to help her find the direction of the force in part A:

So the force on the charged particle would be v cross B . If it's v cross B on the charged particle, then the wire has electrons running through it, so the current's going that way, the electrons are going this way. So V would be this way, crossing B . [Natalie, quote 6]

Natalie knows the form of the equation for magnetic force on a charged particle, but needs to apply this equation to current in a wire. By understanding that a current involves movement of electrons, she is able to coordinate the equation with her physics knowledge to solve for the direction of the force. Natalie is explicitly using physics concepts to aid in the execution of her solution, in contrast to Matt who uses physics concepts to start his thinking about the problem.

The above examples show how students can use their knowledge of physics concepts to help them make sense of the problem or execute their solution. We also observed students using their physics knowledge to help them reason through the relationships relevant to the problem. For example, in the quote below, Hawk tries reasoning through the relationship between the wire and magnetic field:

If the wire is straight perpendicular to the magnetic field, then [the force is] zero. Right? ...it's B and L . If I move L , if L were to be straight up – if I were to be completely perpendicular then it just wouldn't work anymore. [Hawk, quote 11]

Although Hawk does not have the correct model in saying that when the wire is “*straight perpendicular to the magnetic field... the force is zero,*” he uses this knowledge to reason through the problem. “*If I were to be completely perpendicular then it just wouldn't work anymore.*”

The evidence presented above shows how students can use their relevant physics knowledge and understanding to help them solve introductory physics problems. We

found that students can use this tool in a variety of ways in their physics problem-solving, even when they held incorrect ideas about the physics. The case studies presented later in the chapter will explore further how students coordinate (or fail to coordinate) their physics knowledge with other tools in their problem-solving.

Extreme values

Checking one's answer by plugging in extreme values is a practice used by expert physics problem-solvers (Reif & Heller, 1982) and this practice is often suggested in physics textbooks and problem-solving guides (e.g., Lang, 2008; Pelcovits & Farkus, 2008; Tipler & Mosca, 2008). In our interviews, we often observed students checking extreme or prototypical values, such as confirming the magnitude of the force at characteristic rotations or asking when the force would be zero. For example, in the quote below Kelly uses extreme values to check her (incorrect) answer to part B, in which she relates the force to the sine of the angle of rotation.

Here's my B-field; here's my wire. So it's perpendicular. Perpendicular here, perpendicular here. So I'm rotating it... The problem is that, at no point – that can't be – it can't be $[I_0B \sin 60]$ because at no point is the force zero. But it would have to be zero at π over 2. [Kelly, quote 26]

Kelly first moves her index finger on her right hand, which represents the wire, in 90 degree intervals, noting the angle that the 'wire' makes with the 'B-field' (her left hand) at each of these four different rotations. She notes that at each of these rotations the wire and B-field are still perpendicular. She then compares this finding with the equation she is using and reasons that based on this equation, "*[the force] would have to be zero [when I rotate the wire by] π over 2.*" However, she reasons that physically, "*it can't be [this answer] because at no point is the force zero.*" Kelly then goes on to try a different

approach, based on the discrepancy she notes when using these tools, by using extreme values to help compare her mathematical model with her physical model.

We found that students used extreme or prototypical values when checking their answers, but also, more interestingly, to help them as they were planning or executing their solution. For example, in the quote below, Matt is planning his solution to the part E of the problem, preparing to draw a sine curve.

So [the graph] would be sort of a steeper sine curve, and when phi is 90, it'll be zero, and it starts off where the sine of phi is one. So it's actually going to look like this, a cosine curve of sorts. [Matt, quote 18]

While determining the features of the graph, Matt checks the values of the force when phi, the angle of rotation from the starting position, is zero and ninety degrees. After checking the rotation at specific values, Matt draws the appropriate curve.

In our interviews, we found that students made use of extreme values as a key tool in their approach to highly spatial introductory physics problems. The evidence presented above suggests that students check or plug in extreme values in their problem-solving, and that this practice can influence the next steps in their solution. In the case studies section of this chapter, we further explore how students coordinate their use of extreme values with other tools, such as representations or math calculation tools, to solve this highly spatial physics problem without the explicit use of an internal pictorial representation.

Right-hand rule

The right-hand rule is a kinesthetic tool used in mathematics and physics to help characterize the orientation of three perpendicular vectors. The right-hand grip rule

relates an axis to the direction of a rotation about that axis. For the purposes of this dissertation, both tools are labeled under the umbrella of *right-hand rule*. In these interviews, students used the right-hand rule to help them represent and execute a cross-product in an equation, determine the direction of a force of the magnetic field on the wire, and represent the direction of the magnetic field generated from the current in the wire. The quote below illustrates how students typically used this tool in problem 3, part A:

So the direction of the force, I can use the right-hand rule again. So [the current] points up and then curl my hands towards the right. Up and towards the right, so I think, the force would be into the page. [Michelle, quote 2]

Michelle physically used her hands to represent the direction of the current in the wire and the direction of the magnetic field while she says the above quote. Combining these hand motions using the heuristics of the right-hand rule led her to determine the direction of the force. In our interviews, the right-hand rule was always used in combination with students' use of their hands, a physical representation of the current-carrying wire, magnetic field, etc. However, theoretically students could also use the right-hand rule in combination with mental-pictorial representations, forgoing the need for their actual hands.

Although most students only used the right-hand rule in part A of the problem, some students used this tool even on parts of the problem that did not ask for the direction of the force. In the quote below, Jax uses the right-hand rule as she is solving for the magnitude of the force in part B:

What would the magnitude of the force on the wire be? ...B, I cross B. now I is – so it would be, it would still be kind of, it would still be perpendicular to both. Instead of going into the page, it's more downward. [Jax, quote 12]

While reasoning through her solution, Jax uses her hands to represent the rotated wire, B-field, and resulting direction of the force. Based on her hand movements, the right-hand rule appears to scaffold her physical representation while she solves for the magnitude of the force.

This evidence presented above suggests that while students most often used right-hand rule in solving for the direction of the force in part A, some students appeared to employ this tool to gain insight on other parts of the problem as well. The right-hand rule is a well-known heuristic in physics problem-solving, although it can also be used in other disciplines, most notably mathematics. As seen in the next section, other mathematical tools play an important role in student problem-solving in physics.

Math calculation

Mathematics is considered to be the ‘language’ of physics and an integral part of physics problem-solving. In our interviews, we observed students using math-calculation tools, such as algebra, geometry, and trigonometry, to help them execute their solutions to physics problems. As seen the quote below, Maverick uses basic algebra and trigonometry to execute his solution to problem 3 part B:

F equals I times B cross L, magnitude of which is IBL sine theta, which is now 60 instead of 90, so IBL sine 60 degrees. Sine of 60 is radical three over two. IBL radical three over two. So the magnitude of the force per unit length is I times B times radical three over two. [Maverick, quote 4]

After deciding that the angle between the magnetic field and current-carrying wire is 60 degrees, instead of 90 degrees, Maverick uses math-calculation tools to help him execute his solution.

While parts of the problem required students to use these calculation tools to arrive at the correct solution, we also found instances in which the students used mathematical-calculation tools in more sophisticated ways; for example, students used math calculations to help them gain additional insight about the problem and their solutions. For example, in the quote below Hawk uses trigonometry, in addition to other tools such as extreme values and physics reasoning, to help him plan and think through a solution to problem 3 part B.

So as you tilt the wire up, there's less and less force. 'Cause sine of zero is zero. Therefore, [the angle] must be from the normal, right? 'Cause it's supposed to be perpendicular. So, sine of zero is zero. [Hawk, quote 6]

Hawk anchors his thinking on the knowledge that the “*sine of zero is zero.*” His reasoning appears to be guided by this trigonometric ‘tool’, as he tries to figure out what angle to use in his solution.

All students used math-calculation tools to some degree in solving these physics problems. While students often used these tools when required to do so for parts of the problem that required algebraic and/or trigonometric manipulation of the equations, we also found, as seen in the quote above, that students used math-calculation tools to help them better understand the problem and plan their solutions.

Mathematical representation

As seen in the literature on the use of mathematics in physics, there are a number of ways in which students can use mathematical representations in problem-solving, e.g., to help perform calculations, map physical concepts, remember rules or results taught during class, or bridge other mathematical ideas (Bing, 2008). We found evidence of students using mathematical representations in diverse ways in our interviews as well. For part A, most students cited a formula to use in calculations, as illustrated in the quote below:

So since I'm given the long wire, I'm thinking $IL \times B$. Can I have the formula sheet?... Okay, so I'm just going to write that down on the side. [Jax, quote 2]

Jax appears to be relying on the mathematical representation to as a memory tool to help her structure her solution and writes it on the side of her paper for reference. She refers to this equation in other parts of her solution, not necessarily using it to calculate, but as a starting point for her thinking. Similarly, Andrew uses part of the mathematical representation to guide his thinking as he solves for magnitude of the force in part D.

That's simply a function of the angle because the cross product is – the magnitude of the cross product is related to the sine of the angle between the two vectors we're using... [Andrew, quote 7]

Andrew is using the equation to help him understand the relationship between the magnetic force and angle of rotation about the z-axis. By recognizing that the cross product contains a dependence on the angle between two vectors, Andrew is able to use the mathematical representation to proceed in reasoning out his solution.

While the use of a mathematical representation was often productive in guiding student thinking on these physics problems, it was also a crutch for some students, as seen when Green uses the equation to solve part B:

I'm just going to guess and substitute 30 degrees into a formula. I don't know what the question is exactly asking for. [Green, quote 5]

Green uses the equation as a fall-back way to arrive at a solution, using the infamous strategy of plug-and-chug. Without coordinating this tool with any others, she has difficulty arriving at a solution for this problem, but seems to recognize that using mathematical representations in this way is not fruitful: *"I'm just going to guess and try..."*

Based on the evidence presented above, we find that students can use mathematical representations in diverse ways on these highly spatial physics problems. While use of this tool was often productive in guiding student thinking, we also observed students using the equations in ways that were less productive, blindly plugging in values to solve the problem. Mathematical representations were also often used in coordination with other representations of the problem, such as graphical representations, described in the next section.

Graphical representation

For parts C and E, students were asked to draw a graph of the magnitude of the force (per unit length) versus the angle of rotation. To complete this part of the problem, students had to use their understanding of how to represent a physical or mathematical idea graphically, therefore all students used the tool of graphical representation to some

degree. In addition to helping them complete these parts of the problem, drawing the graphs seemed to also help some students structure their thinking:

So let's just draw a nice sine graph and show that this [draws axes and begins drawing sine curve] Oh. That was stupid. I didn't think about where my axes were going because it's actually a cosine graph because when the angle is zero, the magnitude is increased. When the angle is 90, [I get a] magnitude of zero. [Jack, quote 7]

During this quote Jack begins drawing his graph and axes to answer part E of the problem. He starts off thinking that he should draw a sine graph, which, presumably, he gets from the equation. However, the problem asks to graph the magnitude of the force as a function of the angle of rotation, not the angle between the magnetic field and current, and thus the sine curve should be shifted by π over 2. As Jack begins drawing, he notes that his graphical representation does not fit the physical situation. He notes the discrepancy between his mathematical representation (sine of the angle) and the value of the force at 0 and 90 degrees. The act of drawing the graph appears to help Jack realize that he needs to account for the initial position of the wire and start his graph from the maximum value.

While the students were required to use graphical representations in parts C and E, there were a few occasions in which students used this tool to help them better understand the other parts of the problem, as illustrated in the following quote from Michelle when she is working on part B:

Well, I guess the graph might be easier, when I think of the general trend, as it rotates further. Like here, it will hit 90, and here, it will hit 90 again. [uses hand motions and pen to represent the rotation] So will it? 30 degrees from the x-axis... [Michelle, quote 18]

Part B of the problem only asks students to calculate the magnitude of the force when the wire is rotated 30 degrees from the axis. However, Michelle relies on thinking about how to represent the rotation of the wire graphically, by considering a general trend. To get a sense of how the force changes throughout the rotation, Michelle checks the value of the force at different rotations of the wire and finds that the B-field and wire are still perpendicular: “*here, it will hit 90 and here, it will hit 90 again.*” By thinking of how to draw the relationship on a graph, Michelle is prompted to use different tools, such as physical representation and extreme values, to help her think about how to solve for the magnitude of the force.

All students used graphical representations to some degree to answer parts C and E of this problem. Evidence suggests that students used this tool to not only complete their solutions to these parts, but also to mediate and supplement their thinking on other parts of the problem.

Pictorial representation

Every student used a pictorial representation to some degree in their problem-solving. Most students began by using the diagram given to help them understand the problem, as evidenced by the following quote from a student just after reading the problem. “*Okay, so I’m just going to look at the diagram and kind of understand it. [looks at paper]*” [Jax, quote 1] From the video, we can see that Jax spends a moment looking at the pictorial representation before proceeding with the problem. Students also added their own diagrams to help them represent the problem. In the quote below, Gummy indicates that she is using the diagram to help her better understand the directions of the current and magnetic field.

So in a constant B-field. So this is the direction of the current [draws on diagram]. The current is going up, B is going this way [looks at diagram while using hand motions to represent current and B-field]. [Gummy, quote 2]

Looking at the video, we can see that she explicitly says “*so this is the direction of the current*” while drawing on the diagram and further uses the diagram to understand the remaining components of the problem.

In addition to using the pictorial representation(s) to help them understand the problem, students also referred back to the diagram when executing their solutions. Almost every student used a pictorial representation in combination with the right-hand rule and physical representation, to help them find the direction of the force in part A. In Figure 5.6, Jax is using the diagram to help her execute the right-hand rule. As she is using her hands to do the right-hand rule, Jax is looking at and pointing to the diagram on the paper.

In addition, some students looked back to the pictorial representation to help them on later parts of the problem. In the following quote, Michelle refers back to her own diagram and the one at the top of the page while working on part B of the problem.

And then going back to this question: see it from the right. So let's say, I'm looking onto the x-axis. [refers to her earlier drawing and adds to it] And so, it would be B-field coming at me. And then, there is this wire that just got rotated 30 degrees, so it's at an angle. What I want to find that angle between the wire and the magnetic field. [Michelle, quote 16]

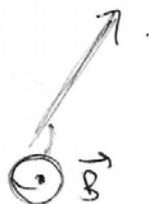
Michelle is looking at the diagram she drew on an earlier attempt to solve part B (Figure 5.7). With the aid of her drawing, she reorients her perspective of the problem so that she is “*looking onto the x-axis*” and considers the directions of the relevant variables: “*B-*

field [is] coming at me... and there is this wire that just got rotated 30 degrees.” Similar to the previous examples, Michelle uses the pictorial representations to help her understand the problem and plan a solution, “*what I want to find is that angle between the wire and the magnetic field.*” She uses both the given diagram and her own drawing to look at the rotation of the wire in the B-field from different angles.

Figure 5.6. Student referring to the pictorial representation drawn on the paper while using her hands to do the right-hand rule.



Figure 5.7. Student diagram rotating the coordinate axes and showing the B-field and rotated wire.



As discussed in these examples, students used pictorial representations throughout the problem-solving process. They used the diagram given, as well as their own drawing, alone and in coordination with other tools. These pictorial representations are limited to the two dimensions offered by the paper; in the next section we'll discuss how students used their hands, pen, and paper for three-dimensional 'pictorial' representations—what we call physical representations.

Physical representation

Most accounts of representation use in physics problem-solving include words, equations, graphs, sketches and diagrams (D. Rosengrant, van Heuvelen, & Etkina, 2006). In addition to these representations, we observed students using their hands, writing instrument, and other physical tools during their problem-solving. Students often used physical representations to help them see different aspects about the problem or manipulate their representations to compare different parts of the problem or check extreme values. For example, in Figure 5.8, Michelle uses her hands to represent the rotation of the wire in the magnetic field. As she makes these motions, she describes her thinking:

So constant B-field this way. This way, it's 90 [degrees], and rotate it 30 degrees this way. So this between here and the z-axis it would be 60 degrees, cause it's in the yz plane. But this angle is a little different – 60 there. So the z... is there, so 60 degrees with the z-axis. But the magnetic field, well the B-field points in the x direction. [Michelle, quote 9]

Michelle uses her hands and pen to see the different angles the wire makes with the axes. She moves her hands and pen to create dynamic pictures of the problem situation in real time. Michelle uses the physical representation much like a three-

dimensional pictorial representation, helping her see the components of the problem in real space.

Figure 5.8. Student using her pen and hands (physical representation) to model the x-axis and rotation about the x-axis.



While Michelle looks explicitly at her physical representation; Figure 5.9 shows a more subtle use of this tool, in which Maverick is rotating his hand but not expressly looking at or manipulating his motions. As he rotates his hands, Maverick states:

About the x-axis... I is always pointing like this, and B is always that way, in which case they're always perpendicular to each other. The magnitude doesn't change. [Maverick, quote 8]

Maverick uses his hands to represent the rotation of the wire, but does not look at his motions; therefore it is not clear exactly how he is using this tool in his problem-solving. He could be using his hands as gestures that reduce his cognitive load by reducing the information he has to process (Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001) or

viewing these movements in his peripheral vision to build a three-dimensional physical representation of the problem.

Figure 5.9. Student using his hands to represent the rotation of the wire, but not looking explicitly at his physical representation.



Mental-pictorial representation

As described in the Methods section, we incorporated aspects of the analytical framework used by Stieff and Raje (2008). We coded use of mental-pictorial representations when students explicitly stated that they were “visualizing” or “trying to picture” the problem or rotation, when they used iconic gestures indicating a use of mental imagery (Feyereisen & Havard, 1999), and/or when there was a gap in students’ reasoning that could not be explained through the use of any other tools the students were using at the time.

We often observed students using mental-pictorial representations to help them understand the problem, specifically for parts B and D, which asked students to rotate the wire about the x-axis and z-axis respectively. For example, in the following quote, Kyle describes using mental-pictorial representations on part D.

I'm just trying to visualize what that would look like, so that's – so this is an xy plane, z is coming out of the page, and rotate about the Z-axis. About the Z-axis... [Kyle, quote 13]

Kyle states explicitly how he manipulated his mental images to understand the rotation. From the video, we can see that he is also using his hands to make small gestures, indicating his use of mental-pictorial representations.

We also observed students using mental-pictorial representations to help them execute their solution when they working on part B, with the wire rotated into the page. For example, in the quote below Maverick describes his thought process while solving part B:

I is always pointing like this, and B is always that way, in which case they're always perpendicular to each other. [uses small gestures indicating use of mental-pictorial representations] The magnitude doesn't change as seen from the right, duh! Okay. So cross this out. The magnitude does not change. Magnitude of F equals I magnitude of B magnitude of L. Yeah. [Maverick, quote 8]

During the quote, we see from the video that Maverick is not explicitly looking at any representation of the problem. His gestures indicate use of mental-pictorial representations and while making these representational gestures he determines that angle between the wire and magnetic field is ninety degrees. Using this information, he is then able to solve that the magnitude of the force is constant.

In these interviews, we observed students constructing and manipulating their mental-pictorial representations to help them understand the problem and execute their solutions. Student use of mental-pictorial representations, specifically as they relate to their score on the MRT, is explored further in the next section and in case studies later in the chapter.

Other parts of the problem/test

We gave students five problems to answer, most of which had multiple parts. We observed students using these other problems or their previous solutions as tools to aid their thinking on their current problem. For example, some students looked to later parts of the problem for ‘clues’ to help them with earlier parts, e.g., “*I was just reading the second part to see if I can get any clues.*” [Kyle, quote 5]

Other students used their previous solutions to other parts of the problems as comparisons for their thinking. In the following quote, Michelle returns to her thinking on part B of the problem, after solving parts D and E.

Am I thinking about this right? So it's this way. Well, I thought about this one. [refers to drawing from part D, uses hand motions and pen to represent the different rotations] I was going this way. Going around like that and think about the x-axis. It's going around, so it's not going around this way. It's going around that way now. [Michelle, quote 17]

Michelle states that she is thinking about how she represented and solved another part to the problem “*Well I thought about this one... going this way.*” With the aid of her physical representations, she compares the rotations, trying to figure out how they are different when they are about different axes. “*It's going around, so it's not going around this way. It's going around that way now.*”

These students are using the exact tools that are recommended as test-taking techniques. Test-taking books and websites even suggest looking at other problems for hints. As this evidence suggests, students use the other parts of the problem as tools in their problem-solving – both as sources for hints and, in a more sophisticated manner, as comparison for their thinking on the current problem. We find additional evidence of the influence of students’ previous experience with physics tests and their test-taking strategies in the next section.

Expectations about physics problems

Students’ epistemological beliefs are known to impact how students learn physics and approach physics problems (e.g., Hammer, 1994, 2000; Lising & Elby, 2005). In our interviews, we found that students voiced expectations that influenced their solutions to these physics problems. For example, in the quote below, Green talks explicitly about her expectations on how to maximize her test score to guide her thinking about the solution.

But I’m just going to substitute 30 degrees again just banking on the possibility that at least either B or D is correct. I think probably one of them or both will be wrong, but I’m just going to take that chance and say it’s the same as B. [Green, quote 11]

Green surmises that her strategy on the problem is flawed, but that she could maximize her points for this problem by putting the same answer for both parts, banking on the idea that one part might have the correct answer. Her expectations about how to improve her score on the test appear to guide her solution, rather than the other tools described here. Similarly, Hawk uses a different test-taking strategy to evaluate his solution to part B.

I really think it’s half, cause odds are that would give us nice numbers. I’m going to star this because I wanna look at this again. [Hawk, quote 6]

Hawk second-guesses his answer, based on his expectations about how physics tests are designed; he eventually changes his answer to match his expectations for physics test answers.

Some students were explicit about how their expectations guided their solving of the problems, while others expressed their expectations without much qualification. In the following quote, Kelly explains her difficulty in solving part B of the problem:

So I feel like I'm stuck because on the one hand, I feel like the force shouldn't be constant in its rotating, but then the math says otherwise. [Kelly, quote 26]

Kelly expresses some expectation about how the force should be changing as it rotates, however her comments about her expectations are limited to: *"I feel like the force shouldn't be constant..."* She doesn't say why she believes the force should change with rotation about the x-axis and offers no support for this answer. Her expectation about the problem is an influential tool; however, because even though *"the math says"* the force is constant, she is reluctant to write down this answer.

Student expectations about the nature of physics problems, tests, and how to maximize their performance on these were often very strong influences on their final solutions. We observed students changing their answers to match their expectations about the nature of physics problems and using their expectations on how get more test points to guide their thinking. In the last quote, Kelly described her thinking about the conflict between her expectations about the physics problem and "the math." In addition to using her expectations about the physics problem, this explicit discussion of her thinking indicates use of metacognition, which we discuss further in the next section.

Metacognition

Metacognitive skills are considered as key components to student learning and problem-solving (e.g., Schoenfeld, 1992; White & Frederiksen, 1998). By monitoring and regulating their thought processes in their problem-solving, students can recognize and build on what they understand and decide to take new approaches when appropriate. We found evidence of students' metacognition throughout these problem-solving interviews. We noted expressions of confidence, such as "*I feel like I got that right.*" [Kelly, quote 22] and "*...that should be good. All right.*" [Matt, quote 19]. Similarly, we observed expressions of confusion or doubt about their understanding, such as "*Okay, so I'm having trouble understanding this question.*" [Green, quote 4] and "*I can't imagine that one.*" [Gummy, quote 6]. Furthermore, some students reflected explicitly on their own thought processes or discussed their understandings of the problem and/or solution, as seen in the quote from Kelly in the previous section.

Some students expressed confidence about their own understanding or solution and then built on that understanding or moved to the next part of the problem. For example, in the quote below, Kelly expresses confidence in her thinking about the problem:

What is the magnitude in force per unit length, using the parameters given? Okay, so we need to write – I need to write I in terms of the speed of... right, right, right. This is pretty simple, should be really simple. Okay, I is equal to charge Q times N... [Kelly, quote 4]

Kelly's expressions of confidence, "*right, right, right. This is pretty simple, should be really simple*" occur as she is planning out her solution to part A. After reflecting on whether her strategy will work, she proceeds with her solution. Similarly, students used

their expressed confusion or doubt about their solution to guide their next steps in their problem-solving, as Jack does in the following quote:

So I think I wasn't thinking very close – very carefully about what I was doing when I was thinking about rotation from the x-axis because since the B-field – or rotation – yeah. [pause] Rotating around the X-axis. The B field's in the direction of the X-axis. So actually, contrary to what I was saying before... [Jack, quote 9]

In this quote, Jack reflects on his own thinking about the problem. “*So I think I wasn't thinking very... carefully...*” He recognizes that his own thinking about the problem was incomplete and then goes back to the problem and changes his solution.

Metacognition can be a powerful tool in physics problem-solving, as we discuss further in the case studies. We observed some students monitoring their own understanding during these interviews and using this metacognition to help them decide on their next steps in their solutions.

In this section of the chapter, we described the tools we observed students using to help them solve highly spatial physics problems. In the next section, we will present analysis showing how many students used these tools a highly-spatial part of the problem which asks students to consider the rotation of the wire into the page, thus requiring a three-dimensional picture to understand the relationship between the magnetic field and current-carrying wire.

5.6.2 Examining and quantifying the tools students used specifically for problem 3, part B

As described in the methods section, part B of problem 3 was of primary interest to us, as it is most efficiently solved by transforming a three-dimensional mental image

so that the wire is rotated into the page. After performing this transformation, one can determine that the angle between the wire and B-field is still ninety degrees. However, we found that students used a diversity of tools on this problem and we document and quantify here what tools we observed students explicitly using on this problem, as well as which students used them.

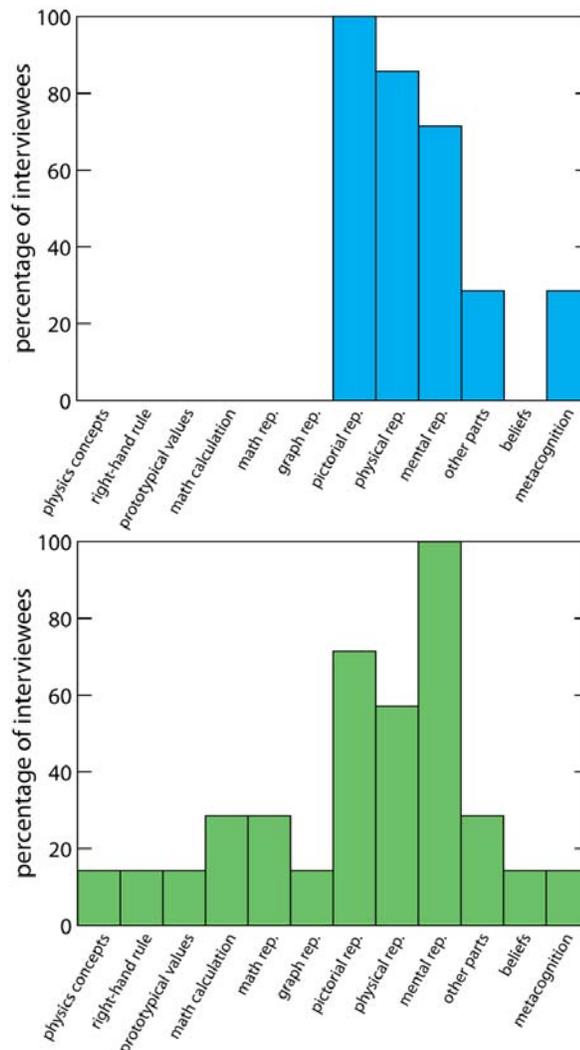
First, we examine how many students answered this part of the problem correctly; these details are shown in Table 5.11. Out of the twelve students we interviewed, seven were able to answer problem 3b correctly. Four out of the seven students who answered the problem correctly scored low on the MRT, all other students scored high. Out of the five students who answered incorrectly, one student scored medium on the MRT, while the other four students scored low. All of the students who were taking the Honors Physics course were able to answer this problem correctly, while three of the Physical Science students answered correctly.

Table 5.11. Number of students answering the problem correctly and incorrectly, separated by course and MRT scores.

	Correct	Incorrect
Physical Science ($N = 8$)	3	5
Honors Physics ($N = 4$)	4	0
Low MRT ($N = 8$)	4	4
Med/High MRT ($N = 4$)	3	1

Next, we look specifically at what tools these students used to answer this part of the problem. In Figure 5.10 we show tools used by students who obtained correct answers, and in Figure 5.11 we show tools used by students who obtained incorrect answers. Within each figure, we break down what tools students used to (a) understand

Figure 5.10. Tool use among students who answered problem 3b correctly to (a) understand the problem and (b) plan and execute a solution.



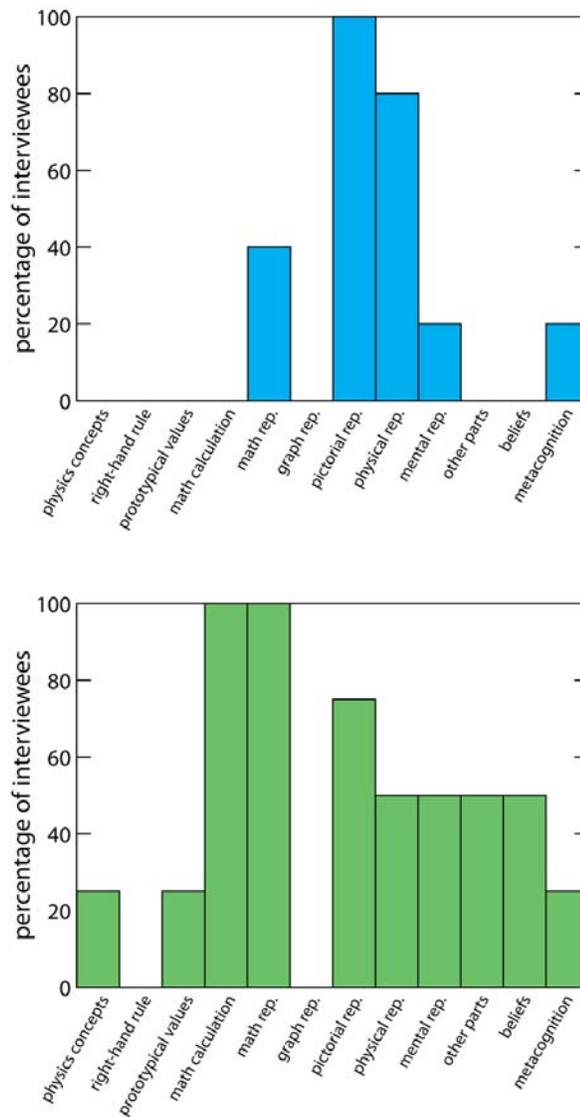
the problem and (b) plan and execute a solution. Thus, in Figures 5.10a and 5.10b we show the tool usage only for students who obtained correct answers. To understand the problem (Figure 5.10a), all students used the pictorial representation, either referring to the picture or drawing their own diagram. Most students also used a physical representation, using their hands and/or pen to represent the rotation. About three-

quarters of students indicated that they used mental-pictorial representations while they were making sense of the problem. To plan and/or execute a solution (Figure 5.10b), all students used mental-pictorial representations to some degree. Nearly three-quarters of the students used pictorial representation to help them execute their solutions, while over half of the students used a physical representation. Two students used a mathematical representation, as well as math-calculation tools and other parts of the problem. One student, Kelly, used physics concepts, the right-hand rule, extreme values, graphical representation, expectations, and metacognition as tools to aid her planning and execution of a solution. This student will be examined in more detail in a case study in the next section. Three students checked their answers, using a mixture of physical, pictorial, and mental-pictorial representations, as well as metacognition.

In Figures 5.11a and 5.11b, we show the tools used by students who answered the problem incorrectly or did not complete the problem. Just like the students who answered the question correctly, all students used the pictorial representation to understand the problem, either referring to the picture or drawing their own diagram. Four out of the five students used physical representations, while only one student used mental-pictorial representations to understand the problem. Two students started with mathematical representations to make sense of the problem and one student used explicit metacognition. After trying to understand the problem, one student chose to move on to the next question, before attempting to plan or execute a solution. All four remaining students with incorrect or incomplete solutions used mathematical representation and calculation as tools to help them plan and/or execute their solutions to problem 3b. Three out of the four students used pictorial representations, while half used physical

representation, mental-pictorial representations, other parts of the problem, and expectations as tools. One student, Hawk, used physics concepts, extreme values, and metacognition to help him plan and execute. Hawk was also the only student with an incorrect answer who attempted to check his answer. He will also be examined in more detail in a case study in the next section of this chapter.

Figure 5.11. Tool use among students who answered problem 3b incorrectly to (a) understand the problem and (b) plan and execute a solution.



We also break down tool usage by pre-course performance on the MRT. In Table 5.12, we display the tools that students with low and medium/high MRT scores used on this problem. Three out of the four students with medium/high MRT scores and six out of the eight students with low MRT scores used mental-pictorial representation as a tool to plan and/or execute their solution. Most students with high MRT scores relied solely on pictorial, physical, and mental-pictorial representations, while students with low MRT scores used a more diverse toolkit in solving the problem. Only one student with a medium/high MRT score used math representation, while six out of eight students with low MRT made use of math representation at some point during the problem. Students with low MRT scores were the only ones to use tools from the categories of *physics tools* and *other tools*.

Table 5.12. Low MRT and high MRT students' tool usage on problem 3b.

	low MRT ($N = 8$)		high MRT ($N = 4$)	
	understanding	executing	understanding	executing
physics concepts	0	2	0	0
right-hand rule	0	1	0	0
special cases	0	2	0	0
math calculation	0	5	0	1
math representation	1	5	1	1
graphical representation	0	1	0	0
pictorial representation	8	5	4	3
physical representation	6	5	4	1
mental-pictorial	4	6	2	3
other parts of problem	2	4	0	0
expectations	0	3	0	0
metacognition	3	2	0	0

In this section, we described the tools we observed students using on problem 3 of our interviews. We showed in detail the tools that students used on part B of the problem, the part of the problem that required the most spatial thinking. Breaking down the tool usage by correctness of solution and scores on MRT, we compared how these tools were used by different subgroups of students. To get a more holistic picture of how individual students coordinated these tools both productively and unproductively, in the next section we present case studies of individual students' solutions.

5.6.3 Case studies from student interviews

In the qualitative results presented thus far, we show that students used certain tools in their problem-solving. However, simply counting events does not give much insight into how these tools were used together. Furthermore, as presented in the previous section, some students were successful at using certain tools, while other students were not. In this section, we present two comparative case study analyses of students to explore the differences between how students used the same tools and obtained different results.

In the first analysis, we present case studies from Maverick and Kyle, both male students in Physical Science. Maverick scored low on the MRT, obtaining a score of 10 out of 30. Kyle scored over a standard deviation higher, obtaining a score 15, which was just above the mean score for Physical Science. Both of these students use mental-pictorial representations, yet Maverick was able to use this tool in his solution successfully, despite his lower MRT score. We first present descriptions of their problem-solving individually; then we compare and contrast their use of different tools in a

comparative analysis, touching on how students use and exhibit facility with tools in one context and fail to use these same tools productively in another.

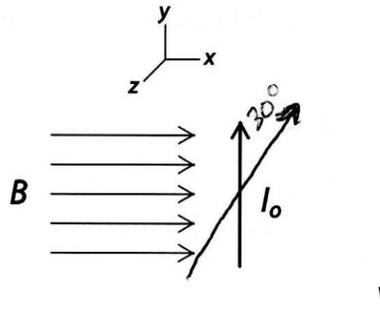
In the second analysis, we examine case studies of Kelly and Hawk, again presenting descriptions of their problem-solving then comparing and contrasting their approaches. In this analysis, neither student relies heavily on the use of mental-pictorial representations. However, Kelly, a female student in Honors Physics, is able to reach the correct solution to part B by coordinating many different tools. Hawk, a male student in Physical Science, also uses these same tools, yet reaches different, and incorrect, conclusions. We explore how Kelly and Hawk use their tools differently and remark on how they determine which tools to use to arrive at their final answer.

Maverick: low MRT score, successful use of mental representations to solve problem

After reading the problem aloud, Maverick immediately uses the diagram, his hands, and the right-hand rule to find the direction of the force on the wire. He finds that the force is into the page, in the negative z direction. Maverick then recalls the relevant equation and uses basic math to arrive at an answer. *“ F equals I times B cross L ... since the angle between the B -field and the current is 90 degrees... I’m going to say F over L equals I times B .”*

Moving on to part B of the problem, Maverick begins to read the question then stops to redraw the diagram with a 30 degree rotation (Figure 5.12) To find the magnitude, he goes back to the equation: *“ F equals I times B cross L , magnitude of which is IBL sine theta, which is now 60 instead of 90, so IBL sine 60 degrees.... So the magnitude of the force per unit length is I times B times radical three over two.”*

Figure 5.12. Maverick's drawing on given diagram representing the rotation of the wire in the xy plane.



To draw a graph for part C, Maverick starts off by drawing and labeling his axes. He then plans his solution: "...it decreases just like a sine function would, I guess. So for small – yeah. So for small changes in the angle, it doesn't change much, but then it changes more quickly as you rotate past 45 degrees." He then labels characteristic rotations on his graph – 90, 180, 270, and 360 degrees.

When Maverick starts reading part D, he pauses to rethink how the wires are rotated in part B. "So I'm a little confused about which direction these axes point in or what it means to be rotated about. I guess it depends on what end you look down the axis from." Using his pen to represent the rotation in the yz plane, Maverick states, "Looking back at part B, the x-axis – I'm thinking now that means the plane perpendicular to... that means rotated in the zy plane instead of what I was originally assuming, which is in the yx plane."

With his new understanding of the problem, Maverick uses small gestures with his hands, the pictorial representation, and mental-pictorial representations to help him solve the problem. "I is always pointing like this, and B is always that way, in which case they're always perpendicular to each other. The magnitude does not change." Figure 5.9

is a screenshot showing his use of small gestures and evidence of mental-pictorial representations in solving this part of the problem.

Going back to part C, Maverick re-draws his axes and quickly determines that the force has a constant magnitude regardless of rotation about the x-axis and draws a straight line for his graph. To finish parts D and E, Maverick uses his previous solutions to parts B and C, checking his thinking about the rotations. “Looking at the page, that’s exactly what I thought. This is the case where F equals $-F$ over L equals I magnitude of B radical three over two. Then I’m going to draw my old graph... OK pretty confident in that. Next question.”

Kyle: medium MRT score, less successful use of mental-pictorial representations

Kyle starts off the part A of problem 3 by going straight to the formula sheet. When coming across the formula $F = qv \times B$, Kyle writes this equation on his sheet, but decides that is the wrong formula and rescans the formula sheet. When he does not find an appropriate equation, he tries to reason physically about the problem: “*Okay, so I know a wire generates a magnetic field, and also that two wires exert a force on each other, so the magnetic field is going this way.*” While reasoning through the physics concepts he knows, Kyle uses his pen in a small gesture to represent the direction of the magnetic field generated by the wire, suggesting use of mental-pictorial representations. He then goes back to trying to find a formula: “*I’m not sure which formula to use for a magnetic field’s force on the wire. It wouldn’t be this one because that’s the magnetic field generated by the positive current... I think it’s the fourth formula...*” He then begins reading part B “*to see if I can get any clues.*” Kyle decides to use the right-hand rule to determine the direction, using $v \times B$ to guide him to $I \times B$. “*So if I used the right-hand*

rule, the I is pointing up, B is pointing to the right, so the force will be into the page. And then the magnitude, in terms of force per unit length... there's no length given, so I'm not sure how to approach that."

Moving onto part B, Kyle reads through the question and models the rotation of the wire by twisting his hands in what appears to be the xy plane. He first approaches the solution using an equation: *"So since I cross B gives you the maximum magnitude of force, rotating it 30 degrees..."* Kyle then draws a quick diagram on his paper, which can be seen in Figure 5.13. His diagram and hand motions indicate that he believes the wire is rotated about the z-axis, in the xy plane. After drawing his diagram, he quickly arrives at his solution: *"I think it's sine of 30 degrees. The sine of 30 degrees is half of what the original magnitude is, so I think the answer to part B is one-half the magnitude on part A because the sine of 30, which is one-half."*

To draw the graph for part C, Kyle starts by drawing his axes and checking the initial value: *"Okay, so from zero to two pi. At zero degrees, it's the max F, max force, so that'll be the F_{max} ."* He then starts reciting the values of various rotations: $\pi/2$, π , $3\pi/2$, and 2π . He consults his diagram while checking the value of the force at $3\pi/2$, saying, *"And then when you go to $3\pi/2$, it would be zero again because then it would be parallel."* He points to his diagram and gestures with his hands to indicate how the wire is flat when rotated by $3\pi/2$. Kyle finishes his curve at 2π , where his value for the force is *"the same as zero [rotation]."* Checking his answer, Kyle checks the value of the force at each of the rotations once more. Using his diagram and gestures, Kyle confirms that *"When you rotate it 90 degrees, then it would be parallel as the B-field, and then that'll be zero, and then when you rotate it down, it will reverse it."*

That'll be negative F_{max} , 3π over 2, it'll be the same thing, it'll be parallel, so it'll be zero, and then 2π will be the same as zero."

Figure 5.13. Kyle's diagram representing the rotation of the wire in the xy plane.



After reading part D of the problem, Kyle spends some time trying to understand the rotation. “Okay, so if that’s rotated about the z-axis, so the wire is originally pointed in the y direction, so when you rotate about the z direction, it’s basically pointing this down.” Kyle uses the picture given in the problem, as well as his hands and pen to determine the rotation. Pausing briefly and using small gestures to indicate the direction of the rotation, Kyle says “I’m just trying to visualize what that would look like, so that’s – so this is an x-y plane, z is coming out of the page and rotate about the z-axis. About the z-axis...” Kyle falls silent briefly and the interviewer prompts him to keep talking. Kyle responds by stating “Okay, so I think the result is the same because when you rotate about the z-axis, the angle between the magnetic field, which is pointing in the x direction, and the wire, which is pointing up in the y direction is still the same... I think that’s the same answer as B, and... the magnitude is one-half the answer in A.” Kyle compares the rotation from part B to the rotation in D, finding that the angle between the wire and the magnetic field does modulate as it is rotated in both situations.

As Kyle has decided that the rotations in part B and D produce similar results, he just copies his solution from part C for part D. “The way I’m looking at it is gonna be the

same answer as C...” He pauses, appearing to double-check his comparison of the two rotations. Small gestures indicate the use of mental-pictorial representations. Appearing satisfied with his reasoning, Kyle fills in the rest of his graph.

Comparative analysis of Maverick and Kyle case studies

Both Maverick and Kyle use a number of different tools to help them solve problem 3. To draw his first graph for part C, Maverick not only considers the graphical representation of the mathematical relationship “*it decreases just like a sine function would,*” but also coordinates this tool with extreme values to obtain the correct curve. When Kyle gets stuck on part A because he does not remember or find the correct equation, he uses other tools, such as physics concepts, mental-pictorial representations, and other parts of the problem to determine which equation to use. However, to solve for the rotation of the wire in the yz plane in part B, Kyle uses pictorial and physical representations to understand the problem, but does not appear to use these tools determine the correct angle between the wire and magnetic field. Instead, he just plugs in the angle given in the problem.

While both Maverick and Kyle first misinterpret rotation about the x -axis as rotation in the xy plane, only Maverick is able to compare the rotations in B and D and see his mistake. Although they obtained different scores on the MRT, both use mental-pictorial representations to help understand the rotation of the wire in the xy plane and solve for the magnitude of the force. Only Maverick, who scored several points lower on the MRT than Kyle, is able to manipulate his mental-pictorial representations to find the correct angle between the wire and magnetic field. Instead, Kyle seems to rely heavily on comparing the rotation from previous parts of the problem.

Finally, neither Maverick nor Kyle spends much time checking or evaluating their answer. Kyle briefly checks while finishing his graph in part C, but uses the exact same tools he used to solve the problem to do so. While Maverick does not check his answers, he does appear to use metacognitive tools to help him compare his thinking between his original thinking on part B and his solution for part D: *“Looking at the page, that’s exactly what I thought.”* He finishes his solution to the problem with an expression of confidence, indicating that he monitored his understanding of the problem and progress to the solution.

The case studies of Maverick and Kyle show how two students with different scores on the MRT use mental-pictorial representations. Maverick uses this tool to help him solve the problem, while Kyle appears to only use the tool to understand the problem, but relies on other tools to execute his solution. In the next section, we discuss Kelly and Hawk, neither of whom use mental-pictorial representations as a predominant tool in their problem-solving.

Kelly: correct solution; diverse tool use

Kelly starts off problem 3 by going straight to the equation. Looking at the formula sheet, she finds the relevant equation, $F = qv \times B$. Using this mathematical representation, she determines the next step. *“Force per unit length, so what we have to do, some sort of interval.”* Pausing briefly, she decides to determine the direction of the force first, using the right-hand rule. While looking at the pictorial representation, she uses her hands to represent the current, B-field, and direction of the force. She correctly determines that the force is into the page, then pauses to check her answer using her

understanding of the physics concepts underlying the equation: “well, if we’re defining positives, it’s the force acting on the positive charges, right?”

Kelly then moves on to determining the magnitude of the force in terms of force per unit length. “Okay, so we need to write – I need to write I in terms of speed of... right, right, right. This is pretty simple, should be really simple.” Confident in her approach, she then uses her knowledge about the definition of current and dimensional analysis to re-write her initial equation, $F = qv \times B$, in terms of current. “So it looks like F_m might be as simple as $I_o B$? F_m per L , that’s qv per L cross v , which is almost I_o ... it’s actually qv over L is I_o divided by N , where N is the number of charge carriers, but that is not something we’re given, so I’m gonna leave the answer as F_m per L is $I_o B$.”

Moving on to part B, Kelly reads the problem aloud and then tries to understand the rotation “So you have this wire, and it’s rotated 30 degrees clockwise, so – I don’t know what they mean ‘as seen from the right’ about the x -axis. I am going to guess that that means if I’m looking down the x -axis, towards the yz plane, then it looks like my wire is at 2 o’clock or something.” While visualizing the rotation, Kelly draws on the diagram and uses her hands to represent the rotation (Figure 5.4). She then returns to the mathematical representation: “So the magnitude is going to change by the function of the angle between the wire and the magnetic field. It’s a cross-product... But now, they’re 30 degrees apart, right? Yeah. So I think the magnitude is going to be $I_o B$... sine 30 degrees. $I_o B$ over 2.” She then begins reading part C and then pauses briefly to check her answer. “I think that’s the right way to think about it.”

Kelly then begins to plan how she’s going to draw her graph, checking for when the force is at a maximum and minimum: “So it’s gonna be a maximum when the angle

between them is 90... when's it gonna be 0?" After asking this question, Kelly returns to her physical representation of the rotation, explicitly representing the B-field, wire, and rotation. *"So this is at 90 degrees. And here – why wouldn't – that can't be right... [maybe I'm] looking at this the wrong way, but there's not gonna be a force when they're parallel. But if I'm thinking about B coming along the x-axis, I must be thinking about it the wrong way. Because they're always gonna be perpendicular."* She re-reads the question, noting that the B-field is pointed in the x-direction in all space. *"So at no point – B cross I, the angle between them... I wonder if it stays constant?"* Pausing briefly, Kelly then tries to understand the equation using her conceptual understanding: *"I'm trying to think about why the magnitude of the force changes as a function of the angle between the magnetic field and the direction that the charges are moving."*

Kelly then moves to part D, where the wire is rotated about the z-axis. Using her hands to represent the rotation, she notes that the force is set at two points. *"[Here] there's going to be no force; it's gonna be a maximum here. This is why, this makes more sense."* After writing her answer for part B for part D, $F = \frac{1}{2} I_0 B$, she begins drawing her graph, using her hands to check the values of the force when the wire is rotated every π over 2 radians. *"Now I'm worried that I'm continuing to think about this wrong... when I rotated π over 2, the force is zero... mathematically, it's either the sine or cosine of the angle. But at π over 2, it's zero, so it's gonna be cosine 30 degrees... so root 3 $I_0 B$ over 2. I feel like I got that right."*

After changing her answer to part D and drawing a cosine curve for part E, Kelly returns to her thinking about part B. *"Here's my B-field, here's my wire. So it's perpendicular. Perpendicular here, perpendicular here. So I'm rotating it – I'm*

wondering whether the force changes at all.” Throughout her reasoning, she is using her hands to represent key angles of rotation of the wire. Not satisfied with her model of a force with constant magnitude, regardless of rotation, she tries to compare this rotation with that in part D. “I am going to be consistent and go with F is equal to I_0B cosine of 30 degrees. But it’s a different – not that one should be preferred over the other, so I’ll say they’re the same.” She writes down her answer, but then goes back to her mathematical reasoning. “So I feel like I’m stuck because on the one hand, I feel like the force shouldn’t be constant in its rotating, but then the math says otherwise. Because if the force were a function of the angle of rotation, at two points it would be zero, but it’s always perpendicular.”

Stuck between her two ideas for a solution, Kelly then tries to think about how to define the B-field differently. “I’m wondering if my problem is that I have to find a B-field, with a vector. I’m supposed to define the B-field with a vector perpendicular.” She then notes the flaws in her reasoning, “if I chose to think about the B-field in terms of a vector, then I would still have the problem...” With this realization, she decides that she can’t figure out how she’s thinking about the problem incorrectly and decides to write down that the magnitude of the force is constant when rotating about the x-axis. “I’m gonna write down F is equal to I_0B as a constant and do the next question.”

Hawk: incorrect solution; diverse tool use

When handed the third problem in the interview, Hawk notes that there are several parts to the problem. “[The problems] are getting longer. I gotta pick my speed up.” After reading the first part to the problem, Hawk immediately begins using the right-

hand rule to find the direction of the force. Looking at the diagram and using his hands to represent the current and B-field, he finds that the force on the wire is into the page.

To find the magnitude of the force, Hawk first asks for the equation sheet. “*So magnitude – I feel like it’s mu naught something. Mu naught four pi r-squared, does that sound right?*” He looks back and forth from the equation sheet and problem, then recalls two equations for force and looks for the variables that match these equations. “*Force equals qvB or ILB . ILB . I know I , yup, I know I . I want force per unit length, so really just F over L equals IB ... let me just be safe with – [force] equals I not magnitude B magnitude sine theta.*” He writes down the full equation for his answer, without substituting for theta.

After reading part B, Hawk first uses his hands and pen to represent the rotation of the wire in the yz plane. He draws on the diagram and begins to think about how the magnitude would change. He starts by thinking about the equation and how to determine the angle between the magnetic field and the wire, “*I feel like it’s sine of theta – it’s not, is it from the normal?*” He then uses his hands to show the rotation of the wire again, “*so as you tilt the wire up, there’s less and less force. ‘Cause sine of zero is zero. Therefore, it must be from the normal, right? ‘Cause it’s supposed to be perpendicular.*” Restating that the sine of zero is zero and using his pen once more to show the rotation, Hawk decides that the force is $I\ell B \sin \theta$, “*even though I really think it’s half, cause odds are that they would give us nice numbers. I’m going to star this because I wanna look again.*”

Moving on to part C, Hawk states, “*I hate questions that build on each other.*” He questions his thinking about the rotation, “*Am I thinking about the axis, right?*” and uses

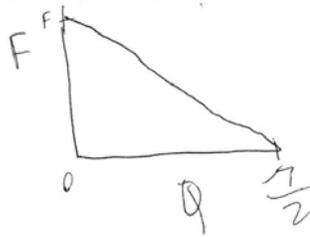
his hands and pen to show different rotations. *“That means if I’m rotating about the... x-axis, that means x should stay the same. It should just move in the y and z.”* Continuing to look at his physical representation of the problem, Hawk again tries to reason through his solution to part B, using math and physics concepts. *“I know that as sine goes to 90 it gets bigger. If the wire is straight perpendicular to the magnetic field, then it’s zero. Right? As the wire – does it even make a difference though? I guess if it’s a wire, a constant wire, right? ...if I were to be completely perpendicular then it just wouldn’t work anymore. It would just move by force.”*

Going back to his reasoning about how to measure the angle, Hawk questions again, *“Is it from the normal or is it not from the normal? Both ways, as your angle goes up...”* He then refers to part D, wondering if it makes any difference which axis you rotate the wire around. *“Alright, we’re going to say that sine theta – why do I feel like theta is from the normal? We have far too many thetas in this class... The wire is flat. All the force is down and the more you tilt the wire...”* After trailing off on this reasoning, Hawk suddenly decides that the angle is thirty degrees. *“We’re going to say that it’s actually going to be a sine of thirty. ‘Cause I feel like it is. So now we’re going to make this one-half I_oB .”* To graph the force as a function of angle of rotation, Hawk reasons that as the angle increases, the force decreases. *“It should be linear, right? We don’t have anything funky going on really.”* Noting the characteristic rotations of 0 and 90, Hawk draws a straight line down from the y-axis to the x-axis (Figure 5.14).

To determine the answers to part D and E, Hawk reasons that *“it shouldn’t make a difference what axis you’re rotated around, it’s always just the angle between the two.”* He puts the same answers from B and C for D and E, noting *“if this assumption is wrong,*

that I made in part B, then all this falls apart, which sucks.” It is unclear from the transcript and video what assumption to which Hawk is referring.

Figure 5.14. Hawk’s graphical representation showing the relationship between magnitude of the force and rotation of the wire.



Comparative analysis of Hawk and Kelly case studies

Both Kelly and Hawk had the most prolific use of tools in their problem-solving out of our interviews; however, Kelly was more successful in using these tools to obtain correct answers. Kelly ultimately arrives at the correct solutions for all parts of the problem, while Hawk is only able to correctly solve for the direction of the force in part A. For example, on parts B and C, both students used almost all of the tools we describe, except neither uses right-hand rule and Hawk doesn’t use metacognition. Despite their use of these similar tools, Kelly eventually responds correctly that the force is unchanged from part A when the wire is rotated about the x-axis, while Hawk incorrectly decides that the force is half of I_0B and that it changes linearly with rotation.

Both Kelly and Hawk were low performers on the MRT, yet both use mental-pictorial representations to some degree in their problem-solving. Kelly first uses mental-pictorial representations to help her understand the problem, when trying to interpret ‘as seen from the right’ quoted in the problem. She also uses it to check her first answer to part B, rethinking the rotation of the wire. To solve the problem, Kelly appears to use

mental-pictorial representations mostly to verify or think about the physics concepts. For example, in part B, at some point Kelly starts thinking about why the force changes as a function of the angle, and appears to use mental-pictorial representation to help her think about the relationship. She does not use mental-pictorial representations heavily in her solutions and, even with the use of physical, pictorial, and mental-pictorial representations, never fully ‘sees’ that the 90 degree angle between the magnetic field and rotated wire in part B. Hawk also uses mental-pictorial representations while he is solving this part of the problem, to determine that “*if I’m rotating... about the x-axis, that means x should stay the same, it should just move in the y and z.*” He also uses mental-pictorial representations to help him compare rotations about the x-axis and z-axis. Ultimately, neither Hawk nor Kelly productively use these representations to arrive at their solutions.

Kelly and Hawk also both use their expectations about physics problems to guide their thinking at some aspect of their problem-solving. Kelly expresses her dilemma with her answer to part B, that the force is constant: “*So I feel like I’m stuck because on the one hand, I feel like the force shouldn’t be constant in its rotating, but then the math says otherwise.*” Although she has reasoned through to the correct solution, Kelly is hesitant to use this answer because her expectation about the problem, that the “*the force shouldn’t be constant in its rotating,*” contradicts her solution. Hawk also uses his expectations to challenge his first solution to part B, claiming that “*I really think it’s half, cause odds are that would give us nice numbers.*” Ultimately, Hawk decides that his answer is one-half I_0B , without using any other tools and claiming that he does this

“[be]cause I feel like it is.” Unlike Hawk, Kelly ultimately decides to go against her gut feeling; using the answer she obtained using other tools.

Although Kelly and Hawk use similar tools in their problem-solving, Kelly coordinates her use of these tools so that they inform how other tools are used and interpreted. For example, in determining the magnitude of the force when the wire is rotated about the x-axis, Kelly combines her physical representation as she uses her hands to represent the rotation, her physics concepts *“there’s not going to be a force when they’re parallel,”* and her math representation *“if the force were a function of the angle of rotation, at two points it would be zero.”* She reaches the correct answer by coordinating these tools together: *“I’m gonna write down that F equals IoB as a constant.”* Hawk, on the other hand, appears to use his tools in isolation. On the same problem, Hawk’s thoughts about the solution seem divorced from one another, e.g. *“I know that as sine goes to 90 it gets bigger”* immediately precedes *“If the wire is straight perpendicular to the magnetic field, then [the force] is zero.”* Although his use of the mathematical representation and calculation tell him that when the angle between the wire and magnetic field is 90, the force is at a maximum, he does not match that to his physics concepts. Ultimately, this lack of coordination between his tools prevents him from arriving at correct, or even consistent, solutions. He decides that as the wire is rotated about either the x- or z-axis, the magnitude of the force is modified by sine of the angle of rotation, yet draws a linear graph to represent this relationship.

Although Kelly and Hawk mostly use the same tools in their problem-solving, Kelly also appears to use metacognition to aid her thinking about the problem. She makes several metacognitive statements throughout her solutions, such as *“I feel like I got that*

right” and *“I don’t know where I’m thinking about this the wrong way.”* Furthermore, she monitors and assesses her progress and thinking on the problem, and uses these reflections to help guide her next steps. For example, as described above, Kelly reaches an impasse after her expectations contradict her solution obtained using multiple representations and extreme values: *“So I feel like I’m stuck because on the one hand, I feel like the force shouldn’t be constant in its rotating, but then the math says otherwise. Because if the force were a function of the angle of rotation, at two points it would be 0, but it’s always perpendicular.”* In this quote, Kelly is explicitly commenting on her thinking and analyzing why she feels ‘stuck.’ After discussing this conflict, she moves to another line of thinking: *“I’m wondering if my problem is that I have to... define the B-field with a vector perpendicular... I don’t recall having to think about it that way.”* Kelly appears to use her metacognition to explore other ways of thinking about the problem, and again to compare these approaches to her prior thinking about these concepts.

5.7 Discussion

The quantitative results of our analysis suggest that there is a slight positive relationship between performance on a test of mental-rotation ability and performance in an introductory physics course, but only in the course for non-majors. Results from student interviews also offer evidence that performance on the MRT is somewhat associated with physics problem-solving in introductory courses. All of the students who scored high, relative to their peers, on the MRT were able to answer problem 3b correctly, while only half of the students who scored low obtained correct answers. Furthermore, all students in

Honors Physics were able to answer problem 3b correctly, regardless of MRT score. However, these conclusions would be stronger if our interview sample contained more Physical Science students with high MRT scores.

Although we found a statistically significant relationship between MRT scores and course achievement in Physical Science, the variability in MRT scores predicted less than 5% of the variability in course achievement. Furthermore, the correlation between the scores in Honors Physics was not significant. The results from student interviews offer insight into the weakness of this relationship. We found that students used several tools when solving a highly spatial introductory physics problem; in this case, facility with mental-pictorial representations only takes students so far. Although most students who solved this problem successfully exhibited some use of mental-pictorial representations, they also constructed their own pictorial, physical, and graphical representations, manipulated a mathematical representations, employed physics concepts and extreme values, and applied the right-hand rule. These results suggest that facility with mental-pictorial representations is useful, but not sufficient to solve even highly spatial physics problems. Our conclusions are limited, however, by the fact that we surveyed and interviewed a group of students at a highly selective research I university. Therefore, these results do not generalize to the entire population of introductory students.

Examination of Kelly's case study also showed that mental-pictorial representations are also not always necessary to solve a highly spatial physics problem correctly. Although Kelly uses mental-pictorial representations to some degree, she ultimately arrives at the correct solution by coordinating several other tools, including

physical and mathematical representations and extreme values. As detailed in the background section, Steiff (2004, 2007) also finds that individuals can use heuristics to solve spatial problems in chemistry, in place of imagistic reasoning. Moreover, experts preferred using analytical tools rather than manipulating mental images (Stieff, 2007). In her work on problem-solving, Adams (2007) found that solving problems requires many different “assets,” which she defines as anything that impacts the problem-solving process, including knowledge, beliefs and cognitive processes. Similar to our results, Adams found that people can use “various combinations of assets in place of others” (p. 301). These findings shed light on why there was a significant relationship between MRT scores and course achievement in Physical Science, but not in Honors Physics. Students in Honors Physics may have more tools at their disposal, with more extensive backgrounds in mathematics and physics. With these additional tools, there may be less of a need to rely on mental-pictorial representations or any other single tool, as Kelly (a student in Honors Physics) exhibited.

Although she found that not all assets are necessary, Adams (2007) noted that weakness in some assets particularly hampered students. She notes that “weaknesses in the beliefs/motivation category can be particularly debilitating,” defining *beliefs* as the ideas that students have about the environment, task, discipline, and their own abilities. In examining Hawk’s case study, we also noted how he used his expectations about physics problems and tests to determine his answer to the problem, even though his use of other tools led him to different answers. The idea that students’ epistemological beliefs have a significant effect on how they approach problems is well-known (e.g., Hammer, 1994), but in our interviews we found that these expectations did more than just shape students’

approach to the problems. We found instances in which students obtained answers that were in direct opposition to their expectations, and they then had to choose which set of tools they should rely on for their final answer.

Deciding which tools to use or, as in the situation described above, which tools to most heavily rely on, requires self-regulation during the problem-solving process. As Schoenfeld (1992) states, “It’s not what you know, it’s how, when, and whether you use it” (p. 60). We coded for metacognition as a tool in student’s problem-solving because of the role it plays in helping to direct student approaches. Schoenfeld (1992) describes the mathematical problem-solving behaviors of students and an expert mathematician on a difficult, unknown, two-part problem. Although the students had more tools (e.g., knowledge of facts and procedures readily accessible), they were unable to solve the problem; the expert eventually arrived at the correct solution. Schoenfeld noted that the students chose an approach quickly and stuck with that decision, even if they were unable to make progress. The mathematician, on the other hand, spent a great deal of time planning his approach, then made constant references about his problem-solving, noting when his approach was working and when he needed to try something new. We observed similar behavior in Kelly, who made many references about her own understanding of the problem and her solution. While we did not note any differences in the use of metacognition between students who answered the problem correctly versus incorrectly, Schoenfeld’s work and Kelly’s case study provide evidence for how metacognition can help individuals productively use and coordinate a diverse set of tools in problem-solving.

We found that different problems and different test settings often resulted in different tool use; context appears to play a role in what tools students decide to use to help them understand and solve problems. For example, in Kyle's case study, we found that although he uses pictorial and physical representations to understand part B of problem 3, he does not seem to use these tools determine the correct angle between the wire and magnetic field. Instead he appears to rely solely on the equation to help him solve the problem. Similarly, students may choose to use mental-pictorial representations and transformations on one problem, such as one on the MRT, and not use this tool on a problem in a different context, such as an introductory physics problem. In comparing the case studies of Kyle and Maverick, we found that MRT scores do not always capture how well students will use mental-pictorial representations on highly spatial physics problems.

The idea that students will use different tools depending on the problem, or how they frame the problem, is not new to physics education research (Hammer, Elby, Scherr, & Redish, 2005). A student's expectations about a situation—in our case a physics problem—shape what they pay attention to, what resources or tools they use, and their actions. For example, Hammer, et al., (2005) describe a discussion of a physics problem among three students. They claim that one student framed the problem as a quantitative one; she cites the different quantities in the problem to apply to a physics formulism. Another student, however, indicates that she views the problem as an opportunity for intuitive sense-making and begins using gestures and physical intuition to answer the question. Each of these students appears to frame the problem differently, resulting in use of different resources. Furthermore, the tools themselves, such as a calculator, can contribute to how students frame a problem by cueing students to perform different

actions, such as calculating (Bing, 2008). In thinking about tests of cognitive ability, specifically tests of mental rotations, these results further suggest that we should be cautious in how we interpret individual student results as they relate to students' use of mental-pictorial representations in physics problem-solving. Students may frame a cognitive test differently than they frame a physics test and therefore use different tools on each.

The challenge in coding for the tools, particularly the cognitive tools, is that we do not have access to what is actually going on in students' heads. We rely on the external indicators—what students speak, look at, and gesture—to best determine what tools they are drawing on to help them solve these problems. In particular, these limitations are especially challenging in determining when students are using mental transformations explicitly. However, the conclusions we draw from this data and analysis do not “live and die” by this coding. As we saw in the case studies, just because a student performs well on the MRT does not mean they will obtain accurate three-dimensional pictures of physics problems, much less be able to solve them correctly. Furthermore, there are plausible ways for students to address highly spatial introductory physics problems without needing to use mental imagery; even if Kelly were using more mental-pictorial representations than we documented, her statements indicated that a student does not need to rely on this tool to complete this problem successfully. Therefore, our conclusion that mental rotations are neither sufficient nor necessary is still viable even with the limitations of what we are able to code for in student problem-solving.

Our results on how students use different tools on highly spatial physics problems provide important insight into how to teach physics problem-solving. Instead of focusing

on developing students' skills with one tool, such as mental-pictorial representations, these results suggest that instructors should encourage students to recognize that there is a diversity of tools they can use to help them solve even highly spatial physics problems. Furthermore, these results have implications on how to address issues of sex differences in the classroom. Although sex differences on tests of mental rotation may relate to differences in course achievement, these results suggest that we shouldn't simply remediate female students to improve their performance on tests of spatial skills. Instead we should encourage all students to practice using and coordinating a variety of tools on different problems. Furthermore, these practices will enable students to progress toward an adaptive expertise in physics, where they can flexibly apply their tools in different contexts (for a review of adaptive expertise, see Bransford, et al., 2000). In the next chapter, we link these results with those from chapters 2 and 3 to further discuss the implications for physics teaching and how these relate to the participation and achievement of under-represented minority students in physics.

Chapter 6

Conclusions and implications for teaching and learning

This thesis covers several issues of relevance for under-represented minority students in introductory physics courses. While racial-achievement differences, gendered responses to interactive teaching methods, and cognitive sex differences on mental rotations seem to be rather disparate issues, in this final chapter we strive to bring together some of the common themes present in each chapter and discuss their relevance for improving and refining physics instruction for under-represented groups. We first recap briefly the conclusions from the main chapters. We then discuss how looking at teaching and learning as apprenticeships may address some of the issues that arose from our work and finally conclude with some thoughts on how these ideas support the importance of discipline-specific education research being conducted within the disciplines themselves.

6.1 Summary of conclusions

6.1.1 Achievement and retention of under-represented minorities in PI courses

In Chapter 2, we revisited the question of whether interactive teaching methods, such as Peer Instruction, can help reduce the gender gap in performance on conceptual surveys and course achievement. The previous work done on this topic is inconclusive. At Harvard University, implementation of fully interactive teaching methods appears to

eliminate gender differences on the FCI in a calculus-based introductory physics course. However, this finding is not replicated at other institutions. Similarly, our analysis of data from lower-level introductory courses at Harvard University finds that the gender gap was persistent even with the use of these teaching methods. Although our analysis demonstrated that the fully interactive teaching methods reduce the gender gap for the most advanced students (perhaps by helping *all* students obtain the maximum score), the results from the lower-level mechanics course and electricity and magnetism courses suggested that more work needs to be done to address the gender gap on conceptual surveys in other courses.

More conclusive are the results obtained concerning racial minorities—Black and Hispanic students—in these courses. Regardless of teaching methods, the under-represented racial- and ethnic-minority students received lower average exam and course grades than their peers, even after accounting for background differences. Interestingly, we did not find these racial differences on the conceptual surveys after controlling for pretest and SAT scores. We hypothesize several explanations for the differing results on the conceptual surveys versus final exams and final grades based on the previous research conducted on these issues. First, we discussed the possibility of stereotype threat effects on the high-stakes exams as opposed to the low-stakes conceptual surveys. We also consider that we were not able to adequately control for differences in mathematics preparation, as the final exams and coursework involved much more quantitative problem-solving than the conceptual surveys. Finally, we discussed that there may be differences between the under-represented minority students and their peers on what it means to study physics, in light of Treisman’s findings at University of California,

Berkeley in calculus courses (Treisman, 1992). From our data and analysis in this dissertation, we are unable to draw any conclusions about the causes for the racial differences we observed, but these results highlight the need for further research on how under-represented racial- and ethnic-minority students experience and learn in reformed physics courses.

The most encouraging results from this chapter were on the relationship between pedagogy and student retention. We found that students who took an interactively-taught introductory physics course were less likely to switch out of a science major than those who took a traditionally-taught lecture-based course by the same instructor. Although our results are limited due to the small sample size of traditionally-taught students and the fact that these students attend a highly-selective university, it is encouraging to think that changes in pedagogy in a single course might impact student persistence in sciences for all students. These results also highlight that much more work is needed on the long-term impacts of teaching in introductory courses.

6.1.2 Student response to interactive teaching: gender differences

In Chapters 3 and 4, we explored student responses to interactive teaching at Harvard University, finding gender differences in student responses to the interactive teaching methods both in surveys and interviews. In a calculus-based introductory electricity and magnetism course at Harvard University, females rated the interactive teaching methods used lower, on average, than males on a post-course survey and student evaluations of teaching. Furthermore, when we controlled for students' expectations about the course at the start of the semester, we found a statistically significant positive

relationship for male students, but not for females. Male students who reported that they expected learning physics to involve active engagement with the material and deep understanding of the concepts tended to report higher ratings of the teaching methods. For female students, the survey data did not explain much of the variability in their ratings of the teaching.

In Chapter 4, we turned to student interviews to better explore why female students rated the teaching methods in the course so low. This data suggested that males and females respond differently to social issues related to the in-class questions and discussion. Furthermore, some students brought up issues related to feeling “tested” when responding to the in-class questions and this concern has meaningful consequences for under-represented students as described in the literature (e.g., stereotype threat). Again, our results are limited due to the unique student population at an elite university; the institutional culture may also be a contributing factor in these student responses. Furthermore, most of the interviewees in our sample were among the best students in the class. However, again, these results point to the need for additional research in how students experience the new teaching methods. While there is substantial video research in PER of students working in small-group tutorials or sections, less data have been collected on what students are doing and saying during collaborative exercises in larger classroom environments.

6.1.3 Sex differences on mental rotation tasks and introductory physics

Finally, we departed from our investigations of interactive teaching methods in Chapter 5, when we examined an often-cited link between sex differences in performance

on spatial ability tests and differences in participation and success in physics. We looked specifically at how performance on mental rotation tests relates to performance in introductory physics and how mental rotation is involved in introductory physics problem-solving. In the Physical Sciences course, we found low, but statistically significant correlations between performance on a test of mental rotations and course performance; however, we did not obtain statistically significant correlations in the Honors Physics course. One student's (Kelly) case study shed some light on the necessity (or lack thereof) of mental rotations in introductory physics problem-solving. Although Kelly did not give evidence that she had an accurate picture of the highly-spatial physics problem, she was able to coordinate other tools to help her solve the problem correctly. As an Honors Physics student, Kelly has more physics and mathematical tools to help her tackle the problem, which may be related to the lack of correlation we obtained in this course by de-emphasizing students' reliance on one tool alone (in this case, mental rotations).

6.2 Insights from an apprenticeship perspective of learning

While each of these results from this dissertation stand alone, it is useful to place their implications for teaching under-represented minority students in a broader context. To do so, we draw from the education literature on cognitive apprenticeship and communities of practice, which provides additional insight on our results and gives suggestions for refining or reforming physics instruction further.

6.2.1 Learning as becoming a part of a community

During Carl Wieman's physics colloquium at Harvard University in April 2009, he talked about how many of his graduate students entered his atomic physics laboratory rather clueless about physics, even after 17 years of success in math and science courses (Wieman, 2009). However, only a few years later, these same graduate students became expert physicists "able to win arguments with me all the time." Wieman claimed that in the research lab, students are "doing the cognitive processes and approach towards thinking and learning that was necessary to make them expert physicists." However, it is not just the cognitive processes that are aiding in the students' transformations, but the fact that these graduate students are apprentices in the lab, being mentored by older students, post-doctoral researchers, and professors to become a part of the community of physicists. These students are learning: (1) how to use the tools—both the physical, experimental tools as well as the more abstract conceptual and methodological tools of the discipline; (2) scientific practices and ways of thinking; and (3) how to speak and argue about physics, even going so far as to be able to win arguments with Nobel prize-winning scientists. These graduate students become enculturated in science, specifically physics, and are accepted gradually within the community as they become senior students and post-doctoral researchers.

A wide body of research in the education community highlights the importance of thinking about learning as becoming a part of a community (Brown, Collins, & Duguid, 1989; Lave & Wenger, 1991; Wenger, 1998). In their book on situated learning, Lave and Wenger (1991) propose this idea of learning as a process of participation in a community of practice, highlighting the apprenticeships of tailors in West Africa. Similar to science

graduate students, tailors' apprentices first start with basic tasks, such as running errands or finishing a garment, and then progress gradually to more sophisticated tasks, such as cutting and piecing. Lave and Wenger argue that trivial activities at the beginning of the apprenticeship take on new meaning when looking at them through the lens of community: “[They] provide a first approximation to an armature of the structure of the community of practice.” (p. 96). As participants, they start on the periphery of the community and become more meaningful, central members gradually.

Unlike physics graduate students, however, the community of research physicists may not be the community in which undergraduate students should strive to participate. However, these students should not be “clueless,” as Carl Wieman put it, about the practices of physicists. Instructors are training introductory students, for the most part, to be scientists, or at least scientifically-minded members of society. Therefore, engaging these students to participate as scientists in the classroom will involve scaffolding and modeling how to use scientific tools and structure scientific arguments, as well as providing opportunities for the students to practice these activities themselves. The class itself will become a local community of scientists, as students and instructors work together to develop scientific ways of thinking and make sense of the physical world.

6.2.1 Applying these ideas to the classroom

How do students learn as apprentices in a classroom? Drawing from the Lave's analysis of tailors, Brown, Collins, and Newman (1989) pose that apprentice students can learn from teachers' modeling, coaching, and fading. During the modeling process, learners observe experts doing the practices of the discipline—such as physics problem-solving.

Through these observations, learners develop a conceptual model of the target task or process. For example, in a physics course, students could observe a scientific refutation to an argument, in which an expert models how to provide evidence to challenge a claim. In coaching, the experts provide support for the learners as they engage in these practices. For example, in mathematical problem-solving, Schoenfeld (1992) asks students three questions as they work in groups: (1) what exactly are you doing?; (2) why are you doing it? And (3) how does it help you? In asking these questions, the instructor is coaching students to monitor their thinking—a behavior found predominantly in expert mathematicians, but rarely in novice students. Finally, as learners become more capable, the experts fade away their participation, giving the learners more opportunities to practice on their own.

After an interview for one of the studies in this dissertation, a student asked: “What is it that physicists do?” Despite successfully completing two full semesters of introductory physics (one taught with traditional lectures, the other using interactive teaching methods), this biology major was leaving physics without much idea about what it means to “do” physics. While this anecdote only speaks to one student’s experiences, it highlights the importance of teaching authentic science practices. The student then asked whether physicists “just sit down with a paper and pencil all day.” While this behavior is typical of some physicists, the problems that practicing theoretical physicists face are far removed from the introductory exercises that students are often asked to complete on homework and exams. Furthermore, the role of experimental work in physics is critical to the field. Although it is not reasonable to expect students to participate in research-level experimental work, the practices of designing experiments, interpreting data and

evaluating methods are still accessible to introductory students (Etkina, et al., 2010). If instructors want students to leave their courses with an understanding of how physicists make sense of the world, the practices that they model and scaffold must be more closely aligned with the community of physicists.

6.2.2 Implications of this perspective for our results on under-represented minorities

While thinking about science learning as an apprenticeship allows for students to engage in more authentic practices, as we have argued in this dissertation, it is important to consider the implications of new teaching ideas and methods for under-represented minorities. In this section, we briefly address how this perspective on teaching and learning addresses our previously presented results on females and under-represented racial and ethnic minorities.

Bridging the achievement gap for under-represented racial minorities (Chapter 2)

In his work at University of California, Berkley, Treisman (1992) found dramatic differences between Black and Chinese students in what it meant to “study math,” as evidenced by the different emphases on time spent going over notes and homework, group learning, and the sense of community when studying the subject. To address these issues, he and others developed programs to combat the isolation of Black and other students and foster a community of mathematics learners. “Our goal was then and continues to be now not merely helping students pass calculus or even to excel at it, but, rather, producing mathematicians.” (p. 368) As described in Chapter 2, the results of this program and others in reducing the racial gaps have been impressive. Given these results and the fact Peer Instruction and other interactive teaching techniques foster group

learning, we might have expected different results on the achievement of under-represented racial and ethnic minorities than what we found. However, the programs described in the literature often go beyond group learning in a single course, emphasizing the importance of a “community life focused on a shared interest in mathematics.” (p. 368)

While the programs that Treisman instilled were beyond the scope of a single course, several of the principles related to the apprenticeship and learning community perspective are relevant for introductory physics. Specifically, Treisman places emphasis on creating a community environment and making the implicit expectations about what it means to study math more explicit. In a PI classroom, for example, more time could be devoted to hearing students’ ideas in whole class discussions. Demaree and Li (2009) report using PI discussions as a time to help students feel like more central participants in the community: “I often share various student ideas with the class to emphasize the idea that multiple explanations can have their strengths in different circumstances, or that multiple representations may all be valid and helpful for specific situations.” (p. 127)

Additionally, as mentioned above, there are implicit expectations about how students should study physics and approach problems. In Treisman’s research, the Chinese students appeared to understand what to do to study math successfully, while the Black students did not. It could be that students from under-represented groups have less outside access to the physics or mathematical communities and therefore fewer opportunities to observe physics or mathematics experts modeling the disciplinary practices. By using classroom time to model and scaffold how to approach problems, make sense of concepts, and learn science, all students, but especially those on the

periphery of the community, will benefit from better understanding and more inclusion in the science community.

Modeling and scaffolding scientific arguments to address the social issues in PI discussions

In Chapter 4, we found that more females than males in our sample reported dealing with social issues related to the discussion parts of PI. As we discussed in the conclusions section of that chapter, resolving these issues in the classroom does not mean removing the discussions. Being able to put forth one's scientific ideas and discuss them with peers is an important skill for all professionals, regardless of discipline, but especially for scientists. Indeed as Lemke (1990) puts in his introduction to his book *Talking Science*, "Learning science means learning to *talk* science... It means learning to communicate in the language of science and act as a member of the community of people who do so." (p. 1) However, as we found in these interviews, students may not know, or have differing ideas about, what it means to discuss science. Students may have different backgrounds and different interpretations about what it means to argue and disagree about scientific concepts. As Lampert, Rittenhouse, and Crumbaugh (1996) state, "Experience of academic argument as an amicable mode of interaction in our culture is rare." Therefore, especially for students with less exposure to academic arguments, additional modeling and scaffolding are needed to apprentice students into physics discourse.

Providing access to the physics “toolbox” so students can find multiple ways of solving problems

The results from the chapter on mental rotations downplay the role of mental rotations in physics problem-solving and highlight the importance of coordinating different tools in physics to approach problems. This idea of a toolbox is not new when thinking about problem-solving. However, as Schoenfeld (1992) states with regard to mathematics:

The tools of mathematics are abstraction, symbolic representation, and symbolic manipulation. However, being trained in the use of these tools no more means that one thinks mathematically than knowing how to use shop tools makes one a craftsman. Learning to think mathematically means (a) developing a mathematical point of view—valuing the processes of mathematization and abstraction and having the predilection to apply them, and (b) developing competence with the tools of the trade and using these tools in the service of the goal of understanding structure—mathematical sense-making. (p. 3-4)

While physics teachers often teach the various conceptual, mathematical, and physical tools—such as energy conservation, Taylor series expansions, or free body diagrams—used by expert physicists when problem-solving, students must not only be able to use these tools, but, as Schoenfeld argues, they must also develop a physics point-of-view to know how and when to apply them. Developing this disciplinary point of view also means becoming a part of a community (Resnick, 1989):

[T]he reconceptualization of thinking and learning that is emerging from the body of recent work on the nature of cognition suggests that becoming a good mathematical problem solver—becoming a good thinker in any domain—may be as much a matter of acquiring the habits and dispositions of interpretation and sense-making as of acquiring any particular set of skills, strategies, or knowledge. If this is so, we may do well to conceive of mathematics education less as an instructional process (in the traditional sense of teaching specific, well-defined skills or items of knowledge), than as a socialization process. (p. 41)

While much of physics instruction is focused on skill development in problem-solving or acquisition of an understanding of the concepts, understanding how to apply this knowledge is perhaps even more important. Apprenticing students to “think like a physicist” will not only lessen the emphasis on one particular tool (such as mental rotations) that some may claim under-represented groups are less skilled in using, but help all students approach problems in a more sophisticated way.

6.3 Reflections on discipline-specific education research

The last section of this dissertation is a reflection on the training of physics education research graduate students from within physics departments. As we have argued in this chapter, thinking about learning as becoming a part of a community affords us a meaningful perspective on teaching, particularly for under-represented students. Thinking about learning in this way also shifts the perspective of the researcher—from concentrating on acquisition of knowledge and skills to how students understand and participate in the physics community. It becomes important to examine how students engage in scientific practices, such as constructing and manipulating representations, using hypothetico-deductive reasoning, and designing experiments (Etkina, et al., 2010). Because these practices are now the focus, it becomes even more valuable for graduate students in physics education to relate to and engage in science and physics communities. Conducting research in PER from within the physics department affords the researcher the opportunity to gain perspective on the discipline by not only taking advanced classes in physics, but by being an active member of the physics community.

References

- AAUW. (1992). *How Schools Shortchange Girls*. Washington, DC: American Association of University Women Educational Foundation.
- Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation Problems in Meta-Analysis. *Review of Educational Research*, 58(2), 151-179.
- Abrami, P. C., D'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology*, 82(2), 219-231.
- Adams, W. K. (2007). *Development of a problem solving evaluation instrument; untangling of specific problem solving assets*. Unpublished doctoral dissertation, University of Colorado at Boulder, Boulder, CO.
- Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics - Physics Education Research*, 2, 010101.
- Aguinis, H., Pierce, C. A., & Culpepper, S. A. (2009). Scale Coarseness as a Methodological Artifact: Correcting Correlation Coefficients Attenuated From Using Coarse Scales. *Organizational Research Methods*, 12(4), 623-652.
- Alexopoulou, E., & Driver, R. (1997). Gender differences in small group discussion in physics. *International Journal of Science Education*, 19(4), 393-406.
- Alibali, M. W. (2005). Gesture in Spatial Cognition: Expressing, Communicating, and Thinking About Spatial Information. *Spatial Cognition & Computation: An Interdisciplinary Journal*, 5(4), 307 - 331.
- Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: integration of uni- and multidimensional models. *Studies in Higher Education*, 30(6), 723 - 748.
- Ark, W. S. (2005). *Comparing mental rotation and feature matching strategies in adults and children with behavioral and neuroimaging techniques*. Unpublished doctoral dissertation, University of California, San Diego.

- Aronson, J. (2002). Stereotype threat: Contending and coping with unnerving expectations. In J. Aronson (Ed.), *Improving Academic Achievement: Impact of psychological factors on education* (pp. 281-304). San Diego, CA: Elsevier Science.
- Baenninger, M., & Newcombe, N. (1989). The Role of Experience in Spatial Test Performance: A Meta-Analysis. *Sex Roles, 20*(5/6), 327-344.
- Baenninger, M., & Newcombe, N. (1995). Environmental input to the development of sex-related differences in spatial and mathematical ability. *Learning and Individual Differences, 7*(4), 363-379.
- Baker, D., & Leary, R. (1995). Letting girls speak out about science. *Journal of Research in Science Teaching, 32*(1), 3-27.
- Beichner, R. J. (2009). An Introduction to Physics Education Research. In C. Henderson & K. A. Harper (Eds.), *Getting Started in PER* Available from <http://www.per-central.org/document/ServeFile.cfm?ID=8806>
- Beichner, R. J., & Saul, J. M. (2003). *Introduction to the SCALE-UP (Student-Centered Activities for Large Enrollment Undergraduate Programs) Project*. Paper presented at the International School of Physics "Enrico Fermi" Course CLVI, Varenna, Italy.
- Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1972). *Differential Aptitude Tests: Space Relations, Form T*. New York: Psychological Corporation.
- Bianchini, J. A. (1997). Where knowledge construction, equity, and context intersect: Student learning of science in small groups. *Journal of Research in Science Teaching, 34*(10), 1039-1065.
- Bigelow, J., Butchart, S., & Handfield, T. (2006). Evaluations of Peer Instruction. Retrieved May, 2008, from <http://arts.monash.edu.au/philosophy/peer-instruction/evaluations/index.php>
- Bigelow, J., Butchart, S., & Handfield, T. (2007). Peer Instruction Question Database. Retrieved May, 2008, from <http://arts.monash.edu.au/philosophy/peer-instruction/database/index.php>
- Bing, T. J. (2008). *An Epistemic Framing Analysis of Upper Level Physics Students' Use of Mathematics*. Unpublished doctoral dissertation, University of Maryland, College Park, College Park, MD.
- Bock, R. D., & Kolakows, D. (1973). Further Evidence of Sex-Linked Major-Gene Influence on Human Spatial Visualizing Ability. *American Journal of Human Genetics, 25*(1), 1-14.

- Bodner, G. M., & Guay, R. B. (1997). The Purdue Visualization of Rotations Test. *The Chemical Educator*, 2(4), 1-17.
- Bodner, G. M., & McMillen, T. L. B. (1986). Cognitive Restructuring as an Early Stage in Problem-Solving. *Journal of Research in Science Teaching*, 23(8), 727-737.
- Bonsangue, M. V., & Drew, D. E. (1995). Increasing minority students' success in calculus. *New Directions for Teaching and Learning*, 1995(61), 23-33.
- Botzer, G., & Reiner, M. (2005). Imagery in Physics Learning - from Physicists' Practice to Naive Students' Understanding. In J. K. Gilbert (Ed.), *Visualization in Science Education* (pp. 147-168). Netherlands: Springer.
- Brahmia, S. (2008). *Improving Learning for Underrepresented Groups in Physics for Engineering Majors*. Paper presented at the Physics Education Research Conference, Edmonton, Canada.
- Brahmia, S., & Etkina, E. (2001a). *Emphasizing the Social Aspects of Learning to Foster Success of Students at Risk*. Paper presented at the 2001 Physics Education Research Conference, Rochester, NY.
- Brahmia, S., & Etkina, E. (2001b). Switching students on to science: An innovative course design for physics students. *Journal of College Science Teaching*, 31(3), 183-188.
- Bransford, J., Brown, A. L., & Cocking, R. R. (2000). *How people learn: brain, mind, experience, and school* (Expanded ed.). Washington, D.C.: National Academy Press.
- Brewe, E., Kramer, L., & O'Brien, G. (2009). Modeling instruction: Positive attitudinal shifts in introductory physics measured with CLASS. *Physical Review Special Topics - Physics Education Research*, 5.
- Broggt, E., Sabers, D., Prather, E. E., Deming, G. L., Hufnagel, B., & Slater, T. F. (2007). Analysis of the Astronomy Diagnostic Test. *Astronomy Education Review*, 6(1), 25-42.
- Brotman, J. S., & Moore, F. M. (2008). Girls and science: A review of four themes in the science education literature. *Journal of Research in Science Teaching*, 45(9), 971-1002.
- Brown, J., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational researcher*, 18(1), 32.
- Browne, K. P. (2002). *A case study of how upper-division physics students use visualization while solving electrostatics problems*. Unpublished doctoral dissertation, Oregon State University.

- Burin, D. I., Delgado, A. R., & Prieto, G. (2000). Solution strategies and gender differences in spatial visualization tasks. *Psicologica*, *21*, 275-286.
- Canadian In-Class Question Database: Psychology. (2005). Retrieved May, 2008, from <http://cinqdb.physics.utoronto.ca/questions/psychology/>
- Carli, L. (2001). Gender and social influence. *Journal of Social Issues: Gender, Hierarchy, and Leadership*, *57*, 725-742.
- Carter, C. S., Larussa, M. A., & Bodner, G. M. (1987). A study of two measures of spatial ability as predictors of success in different levels of general chemistry. *Journal of Research in Science Teaching*, *24*(7), 645-657.
- Chabay, R., & Sherwood, B. (2006). Restructuring the introductory electricity and magnetism course. *American Journal of Physics*, *74*(4), 329-336.
- Chi, M., Glaser, R., & Rees, E. (1982). Expertise in Problem Solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1). Hillsdale, NJ: Erlbaum.
- Clancy, E. (1962). Women and physics. *American Journal of Physics*, *30*, 626.
- Clayson, D. E. (2007). Conceptual and Statistical Problems of Using Between-Class Data in Educational Research. *Journal of Marketing Education*, *29*(1), 34-38.
- Clayson, D. E. (2009). Student Evaluations of Teaching: Are They Related to What Students Learn?: A Meta-Analysis and Review of the Literature. *Journal of Marketing Education*, *31*(1), 16-30.
- Clement, J. (1994). Use of physical intuition and imagistic simulation in expert problem solving *Implicit and explicit knowledge: An educational approach*. (pp. 204-244): Westport, CT, US: Ablex Publishing.
- Coffey, M., & Gibbs, G. (2001). The Evaluation of the Student Evaluation of Educational Quality Questionnaire (SEEQ) in UK Higher Education. *Assessment & Evaluation in Higher Education*, *26*(1), 89-93.
- Cohen, E. G. (1984). Talking and working together: Status, interaction, and learning. *The social context of instruction: Group organization and group processes*, 171-187.
- Cohen, E. G., & Lotan, R. (1995). Producing equal-status interaction in the heterogeneous classroom. *American Educational Research Journal*, *32*(1), 99.
- Cohen, E. G., Lotan, R. A., Scarloss, B. A., & Arellano, A. R. (1999). Complex Instruction: Equity in Cooperative Learning Classrooms. *Theory Into Practice*, *38*(2), 80.

- Cohen, P. A. (1981). Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies. *Review of Educational Research*, 51(3), 281-309.
- Cole, M. (1996). *Cultural psychology : a once and future discipline*. Cambridge, MA: Harvard University Press.
- Cole, M., & Engestrom, Y. (1993). A cultural-historical approach to distributed cognition. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations* (pp. 1-46). Cambridge, UK: Cambridge University Press.
- Cole, M., & Wertsch, J. V. (1996). Beyond the individual-social antinomy in discussions of Piaget and Vygotsky. *Human Development*, 39(5), 250-256.
- Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics*, 73(12), 1172-1182.
- Collins, A., Brown, J., & Newman, S. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Cooper, L. A. (1975). Mental Rotation of Random Two-Dimensional Shapes. *Cognitive Psychology*, 7(1), 20-43.
- Croizet, J.-C., & Claire, T. (1998). Extending the Concept of Stereotype Threat to Social Class: The Intellectual Underperformance of Students from Low Socioeconomic Backgrounds. *Pers Soc Psychol Bull*, 24(6), 588-594.
- Cross, K. P. (1998). Why Learning Communities? Why now? *About Campus*, 3(3), 4-11.
- Crossgrove, K., & Curran, K. L. (2008). Using Clickers in Nonmajors- and Majors-Level Biology Courses: Student Opinion, Learning, and Long-Term Retention of Course Material. *CBE Life Sci Educ*, 7(1), 146-154.
- Crouch, C. H., Fagan, A. P., Callan, J. P., & Mazur, E. (2004). Classroom demonstrations: Learning tools or entertainment? *American Journal of Physics*, 72(6), 835-838.
- Crouch, C. H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics*, 69(9), 970-977.
- Crouch, C. H., Watkins, J., Fagan, A. P., & Mazur, E. (2007). Peer Instruction: Engaging students one-on-one, all at once. In E. F. Redish & P. J. Cooney (Eds.), *Research-Based Reform of University Physics* (Vol. 1). College Park, MD: American Association of Physics Teachers.

- d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, 52(11), 1198-1208.
- Demaree, D., & Li, S. (2009). *Promoting productive communities of practice: an instructor's perspective*. Paper presented at the 2009 Physics Education Research Conference, Ann Arbor, MI.
- diSessa, A. A., Elby, A., & Hammer, D. (2002). J's epistemological stance and strategies. In G. M. Sinatra & P. Pintrich (Eds.), *Intentional Conceptual Change* (pp. 237-290). Mahwah, NJ: Lawrence Erlbaum
- Docktor, J., & Heller, K. (2008). *Gender Differences in Both Force Concept Inventory and Introductory Physics Performance*. Paper presented at the 2008 Physics Education Research Conference, Edmonton, Alberta (Canada).
- Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, 29(5), 611-623.
- Dowell, D. A., & Neal, J. A. (1982). A Selective Review of the Validity of Student Ratings of Teachings. *The Journal of Higher Education*, 53(1), 51-62.
- Draper, S. W., & Brown, M. I. (2004). Increasing interactivity in lectures using an electronic voting system. *Journal of Computer Assisted Learning*, 20(2), 81-94.
- Draper, S. W., Cargill, J., & Cutts, Q. (2002). Electronically enhanced classroom interaction. *Australian Journal of Educational Technology*, 18(1).
- Driver, R., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). *Making Sense of Secondary Science: Research into Children's Ideas*. London: Routledge.
- Dufresne, R., Gerace, W., Leonard, W., Mestre, J., & Wenk, L. (1996). Classtalk: A classroom communication system for active learning. *Journal of Computing in Higher Education*, 7(2), 3-47.
- Edens, K. M. (2008). The Interaction of Pedagogical Approach, Gender, Self-Regulation, and Goal Orientation Using Student Response System Technology. *Journal of Research on Technology in Education*, 41(2), 161-177.
- Elby, A., & Hammer, D. (2001). On the substance of a sophisticated epistemology. *Science Education*, 85(5), 554-567.
- Ellis, A. B., Cappellari, A., Lisensky, G. C., Lorenz, J. K., Meeker, K., Moore, D., et al. (2000). ConcepTests. Retrieved May, 2008, from <http://www.jce.divched.org/JCEDLib/QBank/collection/ConcepTests/>
- Entwistle, N., Hanley, M., & Hounsell, D. (1979). Identifying Distinctive Approaches to Studying. *Higher Education*, 8(4), 365-380.

- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data*. Cambridge, MA: MIT Press.
- Esmonde, I. (2009). Ideas and identities: Supporting equity in cooperative mathematics learning. *Review of Educational Research*, 79(2), 1008.
- Etkina, E., Karelina, A., Ruibal-Villasenor, M., Rosengrant, D., Jordan, R., & Hmelo-Silver, C. (2010). Design and Reflection Help Students Develop Scientific Abilities: Learning in Introductory Physics Laboratories. *Journal of the Learning Sciences*, 19(1), 54-98.
- Etkina, E., & Van Heuvelen, A. (2007). Investigative Science Learning Environment – A Science Process Approach to Learning Physics. In E. F. Redish & P. J. Cooney (Eds.), *Research-Based Reform of University Physics* (Vol. Reviews in PER Vol. 1). College Park, MD: American Association of Physics Teachers.
- Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education*, 4(1), 69-111.
- Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, 30(6), 583-645.
- Feldman, K. A. (2007). Identifying Exemplary Teachers and Teaching: Evidence from Student Ratings. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (pp. 368-395). New York: Agathon Press.
- Feng, J., Spence, I., & Pratt, J. (2007). Playing an Action Video Game Reduces Gender Differences in Spatial Cognition. *Psychological Science*, 18, 850-855.
- Feyereisen, P., & Havard, I. (1999). Mental Imagery and Production of Hand Gestures While Speaking in Younger and Older Adults. *Journal of Nonverbal Behavior*, 23(2), 153-171.
- Freeman, M., & Blayney, P. (2005). *Promoting interactive in-class learning environments: A comparison of an electronic response system with a traditional alternative*. Paper presented at the Innovation for Student Engagement in Economics, Eleventh Australasian Teaching Economics Conference, Sydney, Australia.
- Gardner, H. (1993). *Frames of mind: the theory of multiple intelligences* (2nd ed.). New York: Basic Books.
- Geiser, C., Lehmann, W., & Eid, M. (2006). Separating "Rotators" From "Nonrotators" in the Mental Rotations Test: A Multigroup Latent Class Analysis. *Multivariate Behavioral Research*, 41(3), 261 - 293.

- Glück, J., Machat, R., Jirasko, M., & Rollett, B. (2001). Training-related changes in solution strategy in a spatial test: An application of item response models. *Learning and Individual Differences, 13*(1), 1-22.
- Gokhale, A. A. (1995). Collaborative learning enhances critical thinking. *Journal of Technology Education, 7*, 22.
- Goldin-Meadow, S. (2005). *Hearing gesture: how our hands help us think*. Cambridge, Mass.: Belknap Press of Harvard University Press.
- Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining Math: Gesturing Lightens the Load. *Psychological Science, 12*(6), 516-522.
- Green, P. (2002). *Peer Instruction for Astronomy*. Upper Saddle River, NJ: Prentice Hall.
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64-74.
- Halpern, D. F. (2000). *Sex differences in cognitive abilities* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hammer, D. (1994). Epistemological Beliefs in Introductory Physics. [Article]. *Cognition & Instruction, 12*(2), 151.
- Hammer, D. (2000). Student resources for learning introductory physics. *American Journal of Physics, 68*(S1), S52-S59.
- Hammer, D., & Elby, A. (2002). On the form of a personal epistemology. In B. Hofer & P. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 169-190). Mahway, NJ: Lawrence Erlbaum Associates.
- Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In J. P. Mestre (Ed.), *Transfer of Learning: Research and Perspectives*. Greenwich, CT: Information Age Publishing.
- Hausmann, M., Slabbekoorn, D., Van Goozen, S. H. M., Cohen-Kettenis, P. T., & Güntürkün, O. (2000). Sex hormones affect spatial abilities during the menstrual cycle. *Behavioral Neuroscience, 114*(6), 1245-1250.
- Hazari, Z., Sadler, P. M., & Tai, R. H. (2008). Gender Differences in the High School and Affective Experiences of Introductory College Physics Students. *The Physics Teacher, 46*(7), 423-427.
- 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.

- Heller, P., & Hollabaugh, M. (1992). Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups. *American Journal of Physics*, 60(7), 637-644.
- Heller, P., Keith, R., & Anderson, S. (1992). Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving. *American Journal of Physics*, 60(7).
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30(3), 141-158.
- Hoekstra, A. (2008). Vibrant student voices: exploring effects of the use of clickers in large college courses. [Article]. *Learning, Media, & Technology*, 33(4), 329-341.
- Hofer, B. K., & Pintrich, P. R. (1997). The Development of Epistemological Theories: Beliefs About Knowledge and Knowing and Their Relation to Learning. *Review of Educational Research*, 67(1), 88-140.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.
- Howes, E. V. (1998). Connecting girls and science: A feminist teacher research study of a high school prenatal testing unit. *Journal of Research in Science Teaching*, 35(8), 877-896.
- Hrabowski, F. A., & Maton, K. I. (1995). Enhancing the Success of African-American Students in the Sciences: Freshman year outcomes. *School Science and Mathematics*, 95(1), 19-27.
- Hugdahl, K., Thomsen, T., & Erslund, L. (2006). Sex differences in visuo-spatial processing: An fMRI study of mental rotation. *Neuropsychologia*, 44(9), 1575-1583.
- Hughes-Hallet, D., Gleason, A. M., McCallum, W. G., Flath, D. E., Lock, P. F., Tucker, T. W., et al. (2003). *Calculus: ConcepTests*. New York: John Wiley and Sons.
- Hyde, J. S. (2005). The Gender Similarities Hypothesis. *American Psychologist*, 60(6), 581-592.
- Hyde, J. S. (2007). New Directions in the Study of Gender Similarities and Differences. *Current Directions in Psychological Science*, 16(5), 259-263.
- Isaak, M. I., & Just, M. A. (1995). Constraints on the processing of rolling motion: The curtate cycloid illusion. *Journal of Experimental Psychology: Human Perception and Performance*, 21(6), 1391-1408.
- Ivie, R., & Ray, K. N. (2005). Women in Physics and Astronomy. Retrieved from <http://www.aip.org/statistics/trends/reports/women05.pdf>

- Johnson, R. T., Johnson, D. W., & Stanne, M. B. (1986). Comparison of Computer-Assisted Cooperative, Competitive, and Individualistic Learning. *American Educational Research Journal*, 23(3), 382-392.
- Jones, M. G., Brader-Araje, L., Carboni, L. W., Carter, G., Rua, M. J., Banilower, E., et al. (2000). Tool time: Gender and students' use of tools, control, and authority. *Journal of Research in Science Teaching*, 37(8), 760-783.
- Jordan, K., Wüstenberg, T., Heinze, H.-J., Peters, M., & Jäncke, L. (2002). Women and men exhibit different cortical activation patterns during mental rotation tasks. *Neuropsychologia*, 40(13), 2397-2408.
- Kanamori, N., & Yagi, A. (2005). Amount of priming in the difference of mental transformation. *Psychological Record*, 55(1), 91-101.
- Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student Evaluations of Teaching: The Generalizability of Class Means. *Journal of Educational Measurement*, 13(3), 171-183.
- Kember, D., & Wong, A. (2000). Implications for evaluation from a study of students' perceptions of good and poor teaching. *Higher Education*, 40(1), 69-97.
- Knight, J. K., & Wood, W. B. (2005). Teaching More by Lecturing Less. *Cell Biol Educ*, 4(4), 298-310.
- Knight, R. D. (2007). *Physics for Scientists and Engineers: A strategic approach*. Reading, MA: Addison-Wesley.
- Kohl, P. B. (2007). *Towards an understanding of how students use representations in physics problem solving*. Unpublished doctoral dissertation, University of Colorado at Boulder, Boulder, CO.
- Kolitch, E., & Dean, A. V. (1999). Student Ratings of Instruction in the USA: hidden assumptions and missing conceptions about 'good' teaching. *Studies in Higher Education*, 24(1), 27.
- Kosslyn, S. M. (1994). *Image and brain: the resolution of the imagery debate*. Cambridge, Mass.: MIT Press.
- Kost, L. E., Pollock, S. J., & Finkelstein, N. D. (2007). *Investigating the Source of the Gender Gap in Introductory Physics*. Paper presented at the 2007 Physics Education Research Conference, Greensboro, NJ.
- Kost, L. E., Pollock, S. J., & Finkelstein, N. D. (2009). Characterizing the gender gap in introductory physics. *Physical Review Special Topics - Physics Education Research*, 5.

- Kozhevnikov, M., Hegarty, M., & Mayer, R. (2002). Visual/spatial abilities in problem solving in physics. In M. Anderson, B. Meyer & P. Olivier (Eds.), *Diagrammatic Representation and Reasoning* (pp. 155-173). London: Springer-Verlag.
- Kozhevnikov, M., Motes, M. A., & Hegarty, M. (2007). Spatial Visualization in Physics Problem Solving. *Cognitive Science*, *31*(4), 549-579.
- Kozhevnikov, M., & Thornton, R. (2006). Real-Time Data Display, Spatial Visualization Ability, and Learning Force and Motion Concepts. *Journal of Science Education & Technology*, *15*(1), 111-132.
- Kulik, J. A. (2001). Student Ratings: Validity, Utility, and Controversy. *New Directions for Institutional Research*, *2001*(109), 9-25.
- Labudde, P., Herzog, W., Neuenschwander, M. P., Violi, E., & Gerber, C. (2000). Girls and physics: teaching and learning strategies tested by classroom interventions in grade 11. *International Journal of Science Education*, *22*(2), 143 - 157.
- Lampert, M., Rittenhouse, P., & Crumbaugh, C. (1996). Agreeing to disagree: Developing sociable mathematical discourse. In D. Olson & N. Torrance (Eds.), *The handbook of education and human development: New models of learning, teaching, and schooling* (pp. 731-764). Oxford, UK: Blackwell Press.
- Landis, C. R., Ellis, A. B., Lisensky, G. C., Lorenz, J. K., Meeker, K., & Wamser, C. C. (2001). *Chemistry ConcepTests: A Pathway to Interactive Classrooms*: Prentice Hall.
- Lang, H. (2008). *Head First Physics*. Sebastopol, CA: O'Reilly Media.
- Lasry, N., Mazur, E., & Watkins, J. (2008). Peer instruction: From Harvard to the two-year college. *American Journal of Physics*, *76*(11), 1066-1069.
- Lave, J., & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.
- Laws, P. W. (1991). Calculus-based physics without lectures. *Physics Today*, *44*(12), 24-31.
- Laws, P. W., Rosborough, P. J., & Poodry, F. J. (1999). Women's responses to an activity-based introductory physics program. *American Journal of Physics*, *67*(S1), S32-S37.
- Layne, B. H., Decristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education*, *40*(2), 221-232.
- Lemke, J. L. (1990). *Talking science: language, learning, and values*. Norwood, N.J.: Ablex Pub. Corp.

- Leone, G., Taine, M. C., & Droulez, J. (1993). The influence of long-term practice on mental rotation of 3-D objects. *Cognitive Brain Research*, *1*(4), 241-255.
- Lewis, M. (2009). *The impact of multicultural pedagogy in science education on the achievement, motivation, and perspectives of cultural minority students*. Unpublished Masters thesis, Evergreen State College, Olympia, WA.
- Li, Y., & O'Boyle, M. W. (2008). How sex, native language, and college major relate to the cognitive strategies used during 3-D mental rotation. *Psychological Record*, *58*(2), 287-300.
- Lichtenstein, G., Loshbaugh, H., Claar, B., Bailey, T., & Sheppard, S. (2007). *Should I stay or should go? Engineering students' persistence is based on little experience or data*. Paper presented at the ASEE Annual Conference and Exposition.
- Linn, M. C., & Petersen, A. C. (1985). Emergence and Characterization of Sex Differences in Spatial Ability: A Meta--Analysis. *Child Development*, *56*(6), 1479.
- Lising, L., & Elby, A. (2005). The impact of epistemology on learning: A case study from introductory physics. *American Journal of Physics*, *73*(4), 372-382.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Lord, T., & Nicely, G. (1997). Does spatial aptitude influence science-math subject preferences of children? *Journal of Elementary Science Education*, *9*(2), 67-81.
- Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, *74*(2), 118-122.
- Lucas, A. (2009). Using peer instruction and i-clickers to enhance student participation in Calculus. *PRIMUS*, *19*(3), 219-231.
- Lynch, S., Kuipers, J., Pyke, C., & Szesze, M. (2005). Examining the effects of a highly rated science curriculum unit on diverse students: Results from a planning grant. *Journal of Research in Science Teaching*, *42*(8), 912-946.
- Maloney, D. P., O'Kuma, T. L., Hieggelke, C. J., & Van Heuvelen, A. (2001). Surveying students' conceptual knowledge of electricity and magnetism. *American Journal of Physics*, *69*(S1), S12-S23.
- Manis, J. M., Sloat, B. F., Thomas, N. G., & Davis, D. (1989). *An analysis of factors affecting choices of majors in science, mathematics and engineering at the University of Michigan*. Ann Arbor, MI: University of Michigan Center for Continuing Education of Women.

- Mann, V. A., Sasanuma, S., Sakuma, N., & Masaki, S. (1990). Sex differences in cognitive abilities: A cross-cultural perspective. *Neuropsychologia*, 28(10), 1063-1077.
- Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, 52(1), 77-95.
- Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11(3), 253-388.
- Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, 83(2), 285-296.
- Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology*, 99(4), 775-790.
- Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluation of university teaching: a multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective Teaching in Higher Education: Research and Practice* (pp. 241-320). New York, NY: Agathon Press.
- Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education*, 7(4), 303-314.
- Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, 52(11), 1187-1197.
- Marsh, H. W., & Ware, J. E. (1982). Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. *Journal of Educational Psychology*, 74(1), 126-134.
- Martin, W. (1978). Effects of scaling on the correlation coefficient: Additional considerations. *Journal of marketing research*, 15(2), 304-308.
- Marx, J. D., & Cummings, K. (2007). Normalized change. *American Journal of Physics*, 75(1), 87-91.
- Maton, K. I., Hrabowski, F. A., & Schmitt, C. L. (2000). African American college students excelling in the sciences: College and postcollege outcomes in the Meyerhoff Scholars Program. *Journal of Research in Science Teaching*, 37(7), 629-654.

- Mazur, E. (1997). *Peer Instruction: A user's manual*. Upper Saddle River, N.J.: Prentice Hall.
- Mazur, E. (in print). *Physics principles*. San Francisco, CA: Pearson.
- McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education*, 21(2), 150-158.
- McCaskey, T. (2009). *Comparing and contrasting different methods for probing student epistemology and epistemological development in introductory physics*. Unpublished doctoral dissertation, University of Maryland, College Park, College Park, MD.
- McDermott, L. C., Schaffer, P. S., & UWPERG. (2002). *Tutorials in Introductory Physics*. Upper Saddle River, NJ: Prentice Hall.
- McGee, M. G. (1979). Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin*, 86(5), 889-918.
- McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist*, 52(11), 1218-1225.
- McNeill, D. (2000). *Language and gesture*. Cambridge, UK: Cambridge University Press.
- Meltzer, D. E. (2005). Relation between students' problem-solving performance and representational format. *American Journal of Physics*, 73(5), 463-478.
- Miller, A. I. (1984). *Imagery in scientific thought : creating twentieth century physics*. Boston, MA: Birkhäuser.
- Miller, A. I. (2000). *Insights of genius : imagery and creativity in science and art*. Cambridge, MA: MIT Press.
- Miller, R. L., Santana-Vega, E., & Terrell, M. (2006). Can Good Questions and Peer Discussion Improve Calculus Instruction? *Primus*, 16(3).
- Morling, B., McAuliffe, M., Cohen, L., & DiLorenzo, T. (2008). Efficacy of Personal Response Systems ("Clickers") in Large Introductory Psychology Classes. *Teaching of Psychology*, 35(1).
- Mullen, A., Goyette, K., & Soares, J. (2003). Who goes to graduate school? Social and academic correlates of educational continuation after college. *Sociology of Education*, 76(2), 143-169.
- National Center for Education Statistics (NCES). (2008). *The Condition of Education: 2008*.

- National Science Foundation. (2008). Science and Engineering Indicators. *National Science Foundation Report*. Retrieved from <http://www.nsf.gov/statistics/seind08/>
- National Science Foundation. (2010). Science and Engineering Indicators. *National Science Foundation Report*. Retrieved from <http://www.nsf.gov/statistics/seind10/>
- Nelson, C. E. (1996). Student Diversity Requires Different Approaches To College Teaching, Even in Math and Science. *American Behavioral Scientist*, 40(2), 165-175.
- Nettles, M. T., Thoeny, A. R., & Gosman, E. J. (1986). Comparative and Predictive Analyses of Black and White Students' College Achievement and Experiences. *The Journal of Higher Education*, 57(3), 289-318.
- Nicol, D. J., & Boyle, J. T. (2003). Peer Instruction versus Class-wide Discussion in Large Classes: a comparison of two interaction methods in the wired classroom. *Studies in Higher Education*, 28(4), 457 - 473.
- Novak, G. M., Gavrin, A. D., Christian, W., & Patterson, E. T. (1999). *Just-In-Time Teaching: Blending active learning with web technology*. Upper Saddle River, N.J.: Prentice Hall.
- Novak, G. M., Patterson, E. T., Gavrin, A. D., Christian, W., & Forinash, K. (1999). Just in Time Teaching. *American Journal of Physics*, 67(10), 937-938.
- Nunes, T., Schliemann, A. D., & Carraher, D. W. (1993). *Street mathematics and school mathematics*. Cambridge, UK: Cambridge University Press.
- Olson, S., Fagen, A. P., & National Research Council Board on Life Sciences. (2007). *Understanding interventions that encourage minorities to pursue research careers*. Washington, D.C.: National Academies Press.
- Oswald, D., & Harvey, R. (2000). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology*, 19(4), 338-356.
- Otero, V. (2004). *Cognitive Processes and the Learning of Physics Part II: Mediated Action*. Paper presented at the International School of Physics "Enrico Fermi" Course CLVI, Varenna, Italy.
- Pallrand, G. J., & Seeber, F. (1984). Spatial ability and achievement in introductory physics. *Journal of Research in Science Teaching*, 21(5), 507-516.
- Pelcovits, R. A., & Farkus, J. (2008). *Barron's AP physics C* (2nd ed.). Hauppauge, N.Y.: Barron's Educational Series.
- Peters, C. C., & Van Voorhis, W. R. (1940). *Statistical procedures and their mathematical bases* (1st ed.). New York, NY: McGraw-Hill Book Company, inc.

- Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A Redrawn Vandenberg and Kuse Mental Rotations Test - Different Versions and Factors That Affect Performance. *Brain and Cognition*, 28(1), 39-58.
- Peters, M., Lehmann, W., Takahira, S., Takeuchi, Y., & Jordan, K. (2006). Mental Rotation Test Performance in Four Cross-Cultural Samples (N = 3367): Overall Sex Differences and the Role of Academic Program in Performance. *Cortex*, 42(7), 1005-1014.
- Pollock, S. J., Finkelstein, N. D., & Kost, L. E. (2007). Reducing the gender gap in the physics classroom: How sufficient is interactive engagement? *Physical Review Special Topics - Physics Education Research*, 3.
- Potter, W. H., De Leone, C., Ishikawa, C., Blickenstaff, J., & Hession, P. L. (2001). *Significant reduction in gender grade disparities in a reformed introductory physics course*. Paper presented at the 2001 Physics Education Research Conference, Rochester, NY.
- Pribyl, J. R., & Bodner, G. M. (1987). Spatial ability and its role in organic chemistry: A study of four organic courses. *Journal of Research in Science Teaching*, 24(3), 229-240.
- Purcell, E. M. (1985). *Electricity and magnetism* (2nd ed.). New York: McGraw-Hill.
- Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Studies in Higher Education*, 16(2), 129 - 150.
- Rao, S., & DiCarlo, S. (2000). Peer Instruction Improves Performance on Quizzes. *Advances in Physiology Education*, 24(1).
- Redish, E. F. (2003). *Teaching physics with the physics suite*. Hoboken, NJ: John Wiley & Sons.
- Redish, E. F., Saul, J. M., & Steinberg, R. N. (1998). Student expectations in introductory physics. *American Journal of Physics*, 66(3), 212-224.
- Reif, F., & Heller, J. I. (1982). Knowledge structure and problem solving in physics. *Educational Psychologist*, 17(2), 102 - 127.
- Rescher, B., & Rappelsberger, P. (1999). Gender dependent EEG-changes during a mental rotation task. *International Journal of Psychophysiology*, 33(3), 209-222.
- Resnick, L. (1989). Treating Mathematics as an Ill-Structured Discipline. In R. Charles & E. Silver (Eds.), *The teaching and assessing of mathematical problem-solving*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

- Rice, R. E., & Bunz, U. (2006). Evaluating a Wireless Course Feedback System: The Role of Demographics, Expertise, Fluency, Competency, and Usage. *Simile*, 6(3).
- Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the Reliability of the Difference Score in the Measurement of Change. *Journal of Educational Measurement*, 20(4), 335-343.
- Rosengrant, D., van Heuvelen, A., & Etkina, E. (2006). *Case Study: Students' Use of Multiple Representations in Problem Solving*. Paper presented at the 2006 Physics Education Research Conference, Rochester, NY.
- Rosengrant, D. R. (2007). *Multiple representations and free-body diagrams: Do students benefit from using them?* Unpublished doctoral dissertation, Rutgers, The State University of New Jersey, New Brunswick, NJ.
- Roychoudhury, A., Tippins, D. J., & Nichols, S. E. (1995). Gender-inclusive science teaching: A feminist-constructivist approach. *Journal of Research in Science Teaching*, 32(9), 897-924.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Sabella, M. S., Coble, K., & Bowen, S. P. (2008). *Using the resources of the student at the urban, comprehensive university to develop an effective instructional environment*. Paper presented at the 2008 Physics Education Research Conference, Edmonton, Alberta (Canada).
- Sağlam, M., & Millar, R. (2006). Upper High School Students' Understanding of Electromagnetism. *International Journal of Science Education*, 28(5), 543 - 566.
- Schafer, J. L., & Olsen, M. K. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research*, 33(4), 545.
- Scherr, R., & Hammer, D. (2009). Student behavior and epistemological framing: Examples from collaborative active-learning activities in physics. *Cognition and Instruction*, 27(2), 147-174.
- Schneider, M. B. (2001). Encouragement of women physics majors at Grinnell College: A case study. *The Physics Teacher*, 39(5), 280-282.
- Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics*. (pp. 334-370). New York: Macmillan Publishing.
- Schultz, K. (1991). The Contribution of Solution Strategy to Spatial Performance. *Canadian Journal of Psychology-Revue Canadienne De Psychologie*, 45(4), 474-491.

- Seymour, E., & Hewitt, N. M. (1997). *Talking about leaving : why undergraduates leave the sciences*. Boulder, Colo.: Westview Press.
- Shaffer, P. S., & McDermott, L. C. (1992). Research as a guide for curriculum development: An example from introductory electricity. Part II: Design of instructional strategies. *American Journal of Physics*, *60*(11), 1003-1013.
- Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *Journal of Educational Psychology*, *93*(3), 604-614.
- Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations*. Cambridge, Mass.: MIT Press.
- Shepard, R. N., & Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science*, *171*(3972), 701-703.
- Shore, B. (1996). *Culture in mind : cognition, culture, and the problem of meaning*. New York: Oxford University Press.
- Simkins, S., & Maier, M. H. (Eds.). (2009). *Just-In-Time Teaching: Across the disciplines, across the academy*. Sterling, VA: Stylus Publishing.
- Sokoloff, D. R., & Thornton, R. K. (1997). Using interactive lecture demonstrations to create an active learning environment. *The Physics Teacher*, *35*(6), 340-347.
- Spelke, E. S. (2005). Sex Differences in Intrinsic Aptitude for Mathematics and Science? *American Psychologist*, *60*(9), 950-958.
- Spencer, S., Steele, C., & Quinn, D. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*, 4-28.
- Starting Point, Teaching Entry Level Geoscience: ConcepTest Examples. (2008). Retrieved May, 2008, from <http://serc.carleton.edu/introgeo/interactive/ctestexm.html>
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *J Pers Soc Psychol*, *69*(5), 797-811.
- Stieff, M. (2004). *A localized model of spatial cognition in chemistry*. Unpublished doctoral dissertation, Northwestern University, Evanston, IL.
- Stieff, M. (2007). Mental rotation and diagrammatic reasoning in science. *Learning and Instruction*, *17*(2), 219-234.
- Stieff, M., & Raje, S. (2008). *Expertise & spatial reasoning in advanced scientific problem solving*. Paper presented at the 8th International conference on the learning sciences

- Stodolsky, S. S., Salk, S., & Glaessner, B. (1991). Student Views About Learning Math and Social Studies. *American Educational Research Journal*, 28(1), 89-116.
- Strenta, C. A., Elliott, R., Adair, R., Matier, M., & Scott, J. (1994). Choosing and leaving science in highly selective institutions. *Research in Higher Education*, 35(5), 513-547.
- Taasoobshirazi, G., & Carr, M. (2008). Gender Differences in Science: An Expertise Perspective. *Educational Psychology Review*, 20(2), 149-169.
- Tai, R. H., & Sadler, P. M. (2001). Gender differences in introductory undergraduate physics performance: university physics versus college physics in the USA. *International Journal of Science Education*, 23, 1017-1037.
- Terrell, M. (2005). GoodQuestions Project. Retrieved May, 2008, from <http://www.math.cornell.edu/~GoodQuestions/>
- Thacker, B. (2003). Recent advances in classroom physics. *Reports on progress in physics*, 66, 1833-1864.
- Tipler, P. A., & Mosca, G. (2008). *Physics for scientists and engineers* (6th ed.). New York, NY: W.H. Freeman.
- Tobias, S. (1990). *They're not dumb, they're different: stalking the second tier*. Tucson, AZ: Research Corporation.
- Treisman, U. (1992). Studying Students Studying Calculus: A Look at the Lives of Minority Mathematics Students in College. *The College Mathematics Journal*, 23(5), 362-372.
- Turpen, C., & Finkelstein, N. (submitted). The construction of different classroom norms during Peer Instruction: Students perceive differences. *Physical Review Special Topics - PER*.
- Turpen, C., & Finkelstein, N. D. (2009). Not all interactive engagement is the same: Variations in physics professors' implementation of Peer Instruction. *Physical Review Special Topics - Physics Education Research*, 5, 020101.
- Van Heuvelen, A. (1991). Learning to think like a physicist: A review of research-based instructional strategies. *American Journal of Physics*, 59(10), 891-897.
- Van Heuvelen, A. (2001). Millikan Lecture 1999: The Workplace, Student Minds, and Physics Learning Systems. *American Journal of Physics*, 69(11), 1139-1146.
- Vandenberg, S. G., & Kuse, A. R. (1978). Mental Rotations, a Group Test of 3-Dimensional Spatial Visualization. *Perceptual and Motor Skills*, 47(2), 599-604.

- Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, 117(2), 250-270.
- Vygotsky, L. (1981). The instrumental method in psychology. In J. V. Wertsch (Ed.), *The instrumental method in psychology*. Armonk, NY: M.E. Sharpe.
- Wallace, B., & Hofelich, B. G. (1992). Process generalization and the prediction performance on mental imagery tasks. *Memory and Cognition*, 20(6), 695-704.
- Walton, G., & Spencer, S. (2009). Latent Ability: Grades and Test Scores Systematically Underestimate the Intellectual Ability of Negatively Stereotyped Students. *Psychological Science*, 20(9), 1132-1139.
- Ware, J. E., & Williams, R. G. (1975). The Dr. Fox effect: A study of lecturer effectiveness and ratings of instruction. *Journal of Medical Education*, 50, 149-156.
- Watkins, J., & Mazur, E. (2009). Using JITT with Peer Instruction. In S. Simkins & M. H. Maier (Eds.), *Just in Time Teaching: Across the disciplines, across the academy* (pp. 39-62). Sterling, VA: Stylus Publishing.
- Wayman, J. (2003). *Multiple Imputation for Missing Data: What is it and how can I use it?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Webb, N. (1984). Sex differences in interaction and achievement in cooperative small groups. *Journal of Educational Psychology*, 76(1), 33-44.
- Weinberg, B. A., Fleisher, B. M., & Hashimoto, M. (2007). *Evaluating Methods for Evaluating Instruction: The Case of Higher Education*.
- Weiss, E., Siedentopf, C. M., Hofer, A., Deisenhammer, E. A., Hoptman, M. J., Kremser, C., et al. (2003). Sex differences in brain activation pattern during a visuospatial cognitive task: a functional magnetic resonance imaging study in healthy volunteers. *Neuroscience Letters*, 344(3), 169-172.
- Wells, M., Hestenes, D., & Swackhamer, G. (1995). A modeling method for high school physics instruction. *American Journal of Physics*, 63(7), 606-619.
- Wenger, E. (1998). *Communities of practice : learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.
- Wertsch, J. V. (1998). *Mind as action*. New York: Oxford University Press.
- Wesp, R., Hesse, J., Keutmann, D., & Wheaton, K. (2001). Gestures Maintain Spatial Imagery. *The American Journal of Psychology*, 114(4), 591-600.

- White, B. Y., & Frederiksen, J. R. (1998). Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students. *Cognition & Instruction*, 16(1), 3.
- Wieman, C. E. (2009). When the Virtual World is Better than Reality; Teaching Physics with Interactive Simulations. *Harvard Physics Colloquia*. Retrieved from http://media.physics.harvard.edu/video/index.php?id=COLLOQ_WIEMAN_041309.flv&width=640&height=360
- Willoughby, S. D., & Metz, A. (2009). Exploring gender differences with different gain calculations in astronomy and biology. *American Journal of Physics*, 77(7), 651-657.
- Wilson, K. L., & Lizzio, A. (1997). The development, validation and application of the Course Experience Questionnaire. *Studies in Higher Education*, 22(1), 33.
- Wright, R., Thompson, W., Ganis, G., Newcombe, N., & Kosslyn, S. M. (2008). Training generalized spatial skills. *Psychonomic Bulletin & Review*, 15(4), 763-771.
- Wu, H.-K., & Shah, P. (2004). Exploring visuospatial thinking in chemistry learning. *Science Education*, 88(3), 465-492.
- Wutchana, U., Emarat, N., & Etkina, E. (2009). *Are Students' Responses and Behaviors Consistent?* Paper presented at the 2009 Physics Education Research Conference, Ann Arbor, MI.
- Zohar, A. (2003). Her physics, his physics: gender issues in Israeli advanced placement physics classes. *International Journal of Science Education*, 25, 245-268.