Role of Pre-Course Student Characteristics on Student Learning in Interactive Teaching Environments

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citation</th>
<th>Miller, Kelly Anne. 2015. Role of Pre-Course Student Characteristics on Student Learning in Interactive Teaching Environments. Doctoral dissertation, Harvard University, Graduate School of Arts &amp; Sciences.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:14226079">http://nrs.harvard.edu/urn-3:HUL.InstRepos:14226079</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Role of Pre-Course Student Characteristics on Student Learning in Interactive Teaching Environments

A thesis presented
by
Kelly Anne Miller
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
September 2014
Role of Pre-Course Student Characteristics on Student Learning in Interactive Teaching Environments

Abstract

The goal of this dissertation is to broaden our understanding of interactive teaching strategies, in the context of the introductory physics classroom at the undergraduate level. The dissertation is divided into four main projects, each of which investigates a specific aspect of teaching physics interactively. All four projects look towards improving the effectiveness of interactive teaching by understanding how pre-course student characteristics affect the way students learn interactively.

We first discuss lecture demonstrations in the context of an interactive classroom using Peer Instruction. We study the role of predictions in conceptual learning. We examine how students’ predictions affect what they report having seen during a demonstration. We also examine how student predictions affect what they recall as the outcome of the demonstration at the end of the semester.

We then analyze student response patterns to conceptual questions posed during Peer Instruction. We look at the relationship between a student’s tendency to switch their answer and pre-course student characteristics like science self-efficacy.

Next we elucidate response timing to conceptual questions posed over the course of the semester, in two introductory physics classes taught using Peer Instruction. We look at the relationship between student response times and student characteristics
like pre-course physics knowledge, science self-efficacy and gender. We study response times as a way of gaining insight into students thinking in Peer Instruction environments as well as to improve the implementation of Peer Instruction.

Finally, we present work on the role of NB, an online collaborative textbook annotation tool, in a flipped, project based, physics class. We analyze the relationship between students’ level of online engagement and traditional learning metrics to understand the effectiveness of NB in the context of flipped classrooms. We also report the results of experiments conducted to explore ways to steer discussion forums to produce high-quality learning interactions.
**Contents**

- Title Page ................................................................. i
- Abstract ........................................................................ iii
- Table of Contents ......................................................... v
- Citations to previously published work .......................... viii
- Acknowledgments .......................................................... ix
- Dedication ....................................................................... xi

1 **Introduction** .............................................................. 1
   1.1 Overview ..................................................................... 1
   1.2 Pre-course student characteristics .............................. 2
   1.3 Organization of dissertation ....................................... 3
   1.4 Interactive teaching strategies ................................... 4
   1.5 Interactive teaching versus traditional lecture ........... 6
   1.6 Theories of Learning for interactive teaching .............. 10

2 **The Role of Physics Lecture Demonstrations in Conceptual Learning** 11
   2.1 Introduction .................................................................. 11
      2.1.1 The Effect of Misconceptions on Observation .......... 12
      2.1.2 Misconceptions and Memory ................................. 13
      2.1.3 Implications to Instruction: Conceptual Change ...... 14
      2.1.4 Elicit, Confront, Resolve ..................................... 15
      2.1.5 Interactive Lecture Demonstrations ..................... 16
      2.1.6 Role of Predictions ............................................. 16
      2.1.7 Focus of Study .................................................. 18
   2.2 Methods ................................................................. 19
   2.3 Results and Discussion ............................................. 21
   2.4 Conclusion ............................................................. 26

3 **Response Switching and Self-Efficacy in Peer Instruction Classrooms** 28
   3.1 Introduction ............................................................. 29
| 3.2 Methods                                      | 31 |
| 3.3 Results                                      | 33 |
| 3.3.1 Descriptive Statistics on Switching       | 33 |
| 3.3.2 Normalizing Switching                     | 34 |
| 3.3.3 Switching and Self-efficacy               | 36 |
| 3.3.4 Switching and Gender                      | 41 |
| 3.3.5 Switching and CT Difficulty               | 41 |
| 3.4 Discussion                                  | 45 |
| 3.4.1 Class-wide Switching                      | 45 |
| 3.4.2 Switching and self-efficacy               | 45 |
| 3.4.3 Switching and CT difficulty               | 46 |
| 3.5 Conclusion                                  | 46 |

| 4 Conceptual question response times in Peer Instruction classrooms | 48 |
| 4.1 Introduction                                                 | 48 |
| 4.2 Methods                                                      | 51 |
| 4.3 Results                                                      | 53 |
| 4.4 Discussion                                                   | 66 |
| 4.4.1 Incorrect answers take more time                          | 66 |
| 4.4.2 Timing and ConcepTest Switching                           | 69 |
| 4.4.3 Response times vary according to students’ pre-course characteristics | 69 |
| 4.4.4 CTs Responses and Gender                                 | 70 |
| 4.5 Conclusion                                                   | 71 |

| 5 The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class | 72 |
| 5.1 Introduction                                                 | 72 |
| 5.1.1 Applied Physics 50                                         | 73 |
| 5.1.2 NB                                                         | 77 |
| 5.1.3 NB Assessment                                              | 78 |
| 5.1.4 Final Grade                                                | 83 |
| 5.2 Methods                                                      | 85 |
| 5.2.1 Sectioning                                                 | 85 |
| 5.2.2 Seeding                                                    | 86 |
| 5.2.3 Student Learning Metrics                                   | 88 |
| 5.2.4 Qualitative Annotation Metrics                             | 88 |
| 5.2.5 Qualitative Coding                                         | 90 |
| 5.2.6 Quantitative NB Metrics                                    | 92 |
| 5.3 NB Use                                                       | 93 |
| 5.3.1 Results                                                    | 93 |
| 5.3.2 Discussion                                                 | 106 |
Citations to previously published work

Parts of this dissertation cover research reported in the following articles:


Acknowledgments

Almost five years ago, I left a job teaching high school physics to join the Mazur Group for a three month research fellowship. I had no idea at the time that I would end up staying so long or that this decision would have such an impact on my life. I have had incredible mentors along the circuitous path I have taken to get here. The experience of knowing these people and developing friendships with them, has changed my life forever.

I would like to thank Kelvin Chu, Dave Hammond, Kevork Spartalian, Bob Arns, Malcolm Sanders and the rest of the physics department at the University of Vermont for teaching me physics, fostering my passion for teaching, and providing friendship and support that extended well beyond the three years that I spent in grad school there. I would also like to thank Phil Sadler, Andrew Ho and Dan Koretz for teaching me so much during the years I have spent at Harvard and for providing me with another perspective outside the School of Engineering.

I would especially like to thank Eric Mazur for allowing me to join his group in the first place and for taking me on as a student. I feel so lucky to have you as an advisor. Thank you for believing in me and for the countless hours you have spent mentoring me on everything from teaching to physics to life in general.

I am also really grateful to the education subgroup over the years; Nathaniel Lasry, Julie Schell, Laura Tucker, Brian Lukoff, Marinna Madrid, Jonathan Guillemette, Ives Araujo, Jason Dowd, James Fraser, Junehee Yoo and Fauzy Wan. I have learned so much from all of you and have enjoyed working together tremendously. Julie and Nathaniel I know that I would not be here if it were not for the two of you and I am truly grateful.
Acknowledgments

I would also really like to thank the entire Mazur Group for the friendship, camaraderie and adventures. Ben Franta, Orad Reshef and Alex Raymond, I owe you guys a life debt. Sally Kang and Phil Munoz, thanks for being such amazing friends, for teaching me so much physics and for helping me survive all my classes and AP50. To the rest of the group, present and past, I have loved working and hanging out with all of you and hope we continue to have the opportunity to do so in the future. Yu-Ting Lin, Renee Sher, Jon Bradley, Daryl Vulis, Michael Moebius, Sarah Griesse-Nascimento, Nabiha Saklayen, Hemi Gandhi, Yang Li, Edwin Marti and Thierry Sarnet thank you for your friendship and company. A special thanks also to Virginia for always taking care of the details and for being so funny.

Lastly, I would like to thank my parents for providing me with such amazing support and the opportunity to pursue my dreams. Without you, none of this would have been possible and I can’t thank you enough for everything you have done for me.
To my parents, who taught me the power of determination and laughter
Chapter 1

Introduction

1.1 Overview

Learning science has never been more important. Understanding the physical world around us is important as technologies advance, and the job market is demanding a work force that is both technically skilled and science literate. Improving the way that science is taught is critical as we face a shortfall of skilled workers in science-related disciplines. Economic projections indicate that, over the next decade, the United States will need 1 million more science, technology, engineering and mathematics professional than it can produce at the current rate. Currently, only 40% of students who enroll in college with the intention of majoring in a science-related field, actually graduate with a science degree [1]. Clearly, as a nation, we need to start teaching science more effectively if we are going to be able to meet the work force needs of the future. Research has shown that students taught interactively have better conceptual understanding, superior problem-solving skills and higher re-
tention in the sciences than students taught with the traditionally passive lecture. The proven efficacy of these teaching strategies is making it increasingly difficult to justify teaching science using old, ineffectual techniques.

The goal of this dissertation is to broaden our understanding of interactive teaching strategies, in the context of the introductory physics classroom at the undergraduate level. The dissertation is divided into four main projects, each of which investigates a specific aspect of teaching physics interactively. All four projects look towards improving the effectiveness of interactive teaching by understanding how student characteristics affect the way they learn interactively.

1.2 Pre-course student characteristics

Each of the projects discussed in this dissertation involves understanding how student characteristics affect the manner in which they learn interactively. We consider a range of student characteristics, including pre-course academic self-efficacy, pre-course physics knowledge, self-reported gender, engagement and conceptual understanding of a specific topic.

Measurement of student characteristics generally varies from project to project. However, we use academic self-efficacy and pre-course physics knowledge consistently across multiple projects. Self-efficacy refers to an individual’s belief that s/he can successfully complete a task [2]. Self-efficacy is a strong predictor for performance in science courses [3,6] as well as science career choices [4,5,7,8]. Self-efficacy also influences a number of factors that are relevant to learning in a Peer Instruction environment, such as perseverance and self-regulated learning [9].
Chapter 1: Introduction

At the beginning of each semester, we administer a pre-course survey designed to measure students’ belief in their academic ability, specifically in the domain of physics. Details about this survey and self-efficacy are discussed in the Methods section of chapter 3. We measure pre-course physics knowledge using validated conceptual surveys. At the beginning of each semester, we administer a conceptual survey as a pre-test. As the spring semester covers electricity and magnetism, we use the Conceptual Survey on Electricity and Magnetism (CSEM) \[10\]. The fall course covers mechanics and we use the Force Concept Inventory (FCI) \[11\]. Both surveys are validated instruments widely used to measure students’ conceptual understanding of Newtonian dynamics (in the case of the FCI) and basic principles of electricity and magnetism (in the case of the CSEM).

1.3 Organization of dissertation

In chapter 2, we study lecture demonstrations in the context of an interactive classroom using Peer Instruction. We investigate the role of students’ predictions and prior conceptual understanding in their ability to correctly observe and remember physics demonstrations shown to them in class. We measure students’ understanding of the specific concept underlying a lecture demonstration by asking a series of conceptual questions. We also ask students to predict the outcome of the demo. We study the relationships between students’ level of understanding of the concept, their predictions of the demonstration and their ability to observe and remember the demonstration accurately.

In chapter 3, we analyze student response patterns to conceptual questions posed
Chapter 1: Introduction

during Peer Instruction. We look at the relationship between a students’ tendency to switch their answer and student pre-course science self-efficacy, pre-course physics knowledge and gender.

In chapter 4, we analyze response timing to conceptual questions posed over the course of the semester, in two introductory physics classes taught using Peer Instruction. We look at the relationship between student response times and student characteristics such as pre-course physics knowledge, science self-efficacy and gender.

We both chapters 3 and 4, we study the relationship between student characteristics and response timing and switching patterns as a way of gaining insight into student thinking in Peer Instruction environments as well as to improve the implementation of interactive teaching.

In chapter 5, we present work on the role of NB, an online collaborative textbook annotation tool, in a flipped, project based, physics class. We analyze the relationship between students’ level of online engagement and traditional learning metrics to understand the effectiveness of NB in the context of flipped classrooms. We also report the results of experiments conducted to explore ways to steer discussion forums to produce high-quality learning interactions.

In chapter 6, we summarize the work in this dissertation and suggest implications for interactive instruction.

1.4 Interactive teaching strategies

“Interactive teaching” has come to include a long-list of methodologies all involve students learning concepts by actively engaging with the instructor and one another.
Chapter 1: Introduction

Interactive teaching methodologies are student-centric (as opposed to teacher-centric) and push students to construct their own understanding of the material through the process of actively asking and answering questions and engaging with one another.

The four projects presented in this dissertation are based on data collected in five introductory physics classrooms, all of which used one or many forms of interactive teaching. The first three projects were based on classrooms taught using Peer Instruction. Peer Instruction is a well-known student-centered teaching method that engages students during class through a sequence of questioning and discussion. Chapters 2-4 provide a more thorough discussion of Peer Instruction and explore ways in which its implementation could be improved.

Chapter 5 discusses work conducted in a flipped, introductory physics classroom taught using a combination of Project Based and Team Based Learning. Flipped classrooms are designed to be highly interactive. Students learn on their own before coming to class and then, in class engage in activities to help them understand the more difficult concepts. This is opposite to the way classes are traditionally taught. Traditionally, students’ first exposure to a topic is in lecture where they sit and listen passively. They are then left to grapple with the difficult concepts later, outside of class. Team Based and project based learning are two interactive teaching strategies that work well in flipped classrooms and will be discussed in further detail in chapter 5.
1.5 Interactive teaching versus traditional lecture

The majority of instructors believe that the best method for training scientists is the traditional lecture. In 2011, the Mathematical Association of America surveyed over 700 calculus instructors on what they believed was the most effective way to teach. Two-thirds of those surveyed agree with the statement: “Calculus students learn best from lectures, provided they are clear and well-prepared” [12]. There is a common belief in teaching that, because the instructor is the expert in the room, it is his/her job to take the lead role in guiding students through the material. Instructors with this traditional view would argue that because students are novices, active engagement techniques where students are ‘teaching’ each other could only lead to an increase in confusion about the concepts.

In addition to serving as the expert, most instructors also believe that their role is to meet their students’ expectations in the classroom. Students can be resistant, at least at first, to new instructional techniques. If students want informative and well-structured lectures, then the traditional lecture meets students’ expectations by allowing them to passively sit, listen, and take notes. To summarize, even though both students and instructors have considerable experience in teaching and learning, this does not directly translate to them being experts in the assessment of various teaching practices. Science education literature has dozens of studies that quantitatively compare student learning in traditional classrooms to learning in classrooms that employ interactive teaching methods. The bulk of this research indicates that students taught interactively have considerably better conceptual understanding and problem-solving skills than students taught with lecture [13–16]. Furthermore, ac-
tive learning correlates with better attendance, improved student ‘affect’ and higher retention in the sciences [17–19].

One of the most highly cited studies to compare student conceptual learning in traditionally taught environments to interactive classrooms was a meta-analysis conducted by [14]. Hake used six-thousand students’ normalized gain scores on the Force Concept Inventory to compare the effectiveness of traditional instruction and of interactive engagement techniques. He found that the interactive courses are, on average, more than twice as effective in producing basic conceptual understanding as traditional courses. The fourteen traditional courses in Hake’s study, containing a total of 2084 students, achieve a normalized gain of 0.23 +/- 0.04 compared to the 48 interactive engagement classes, containing a total of 4458 students, which achieve an average gain of 0.48 +/- 0.14 [14]. Crouch and Mazur found very similar differences in conceptual learning between students taught traditionally and those taught using Peer Instruction [15]. In another study, Meltzer investigated the effectiveness of an active learning strategy over a seven-year period at four large universities. Meltzer found that student conceptual gains, measured with the Conceptual Survey in Electricity and Magnetism, are three times greater for students in an interactive lecture than the average gain from a national sample of classes that used a traditional lecture strategy [16]. Alongside such conceptual gains, the literature also indicates that, in spite of a shift away from explicit problem solving in research-based instructional strategies, improvements in students’ ability to solve problems are at least comparable to those achieved through traditional lecture.

Physics education literature also reveals that students taught interactively have
Chapter 1: Introduction

a higher level of engagement and are more likely to persist in science-related fields than those taught traditionally. Recently, the Carl Wieman Science Education Initiative conducted a controlled experiment comparing two sections of a class taught using different strategies [19]. During the first half of the semester, both sections were taught using traditional lectures. Halfway through the semester, the teaching strategy for one of the sections was switched to a design informed by ‘deliberate practice’ [20] involving many aspects of interactive engagement for a period of one week. Students’ attitudes toward science and level of engagement were measured using class attendance rates and the Colorado Learning Attitudes about Science Survey (CLASS) [21]. Trained observers also watched both sections to measure student engagement. Before the experiment (halfway through the semester), both sections were statistically equivalent with respect to knowledge, attendance, and level of engagement. At the end of the one week test period, results showed that while engagement and attendance for the control section remained unchanged, in the interactive section, student engagement doubled and attendance increased by 20%.

Other studies [17,21] have demonstrated that students’ attitudes about physics and physics learning improve in introductory courses using Modeling [22]. The Modeling Method addresses many of the drawbacks of the traditional lecture by engaging students in activities that help them organize concepts into viable scientific models. Most importantly, this improvement in student attitude has been found to influence career choices in science-related fields. Watkins found a positive relationship between interactive teaching and persistence in the sciences [18]. Regardless of the race or gender of the student, Watkins found that those enrolled in just one interactively
taught introductory physics class are twice as likely to pursue science as their major compared to those enrolled in only traditionally taught classes. Given the need for STEM-literate professionals discussed in the introduction of this paper, the positive correlation between interactive teaching and student persistence could be considered even more important than conceptual understanding gains. Despite many instructors’ belief that students learn best by listening to expert lectures, the literature demonstrates that this is not the case. The evidence for significantly higher conceptual understanding gains in active learning environments over passive ones is hard to overlook.

Anecdotally, one can imagine why Peer Instruction and other interactive engagement strategies are so effective. As novices, students speak the same language, while most instructors, as experts, have long since lost the ability to speak like a novice. As a result, fellow students can play an essential teaching role in partnership with instructors. Furthermore, students are more likely to be critical of explanations provided by peers than those provided by faculty who they perceive as experts. This critical listening promotes learning and helps novices transition to more expert ways of thinking. In fact, a large body of data indicates that students taught interactively are more engaged and persistent in the pursuit of further study in science-related fields. These results show that, on average, active learning strategies strongly benefit students in both the short and long term.
1.6 Theories of Learning for interactive teaching

It is difficult to study teaching strategies without a discussion of how students learn. Here we provide a brief overview of some of the theories of learning that relate to the chapters that follow. A more extensive discussion of the theoretical frameworks that apply to each of the projects is provided at the beginning of each chapter.

Constructivism is an educational theory that emphasizes the role of experience and existing knowledge to build new knowledge. Constructivists believe that people create understanding through their own experiences and that all human perception is influenced by existing knowledge structures [23]. Social constructivism is the belief that knowledge is built through social interactions of shared learning experiences. The learning process is not a solitary exploration but a process whereby students are constructing knowledge collectively and interactively via a community of learners [24]. The Zone of Proximal Development refers to the difference between what a learner can do on their own and what a learner can do with help. Peer learners are more likely to be at a similar level of understanding than instructors who are experts and forget what it was like to learn the material for the first time. Social interactions with other learners who have similar “Zones of Proximal Development” often push students to the next level of understanding more effectively than interactions with instructors whose understanding of the material often makes it difficult to meet students at their level.

Constructivism, social constructivism and this idea of the Zone of Proximal Development are the underlying philosophies of Peer Instruction and other interactive teaching strategies.
Chapter 2

The Role of Physics Lecture Demonstrations in Conceptual Learning

2.1 Introduction

Lecture demonstrations are an integral part of introductory physics courses. Instructors see demonstrations as a way to help students develop an intuitive understanding of the world and remember concepts [25,27]. Instructors often view demonstrations as a way to liven up lectures [27,28]. However, demonstrations may not be effective in promoting student learning and may even be counterproductive [29,31]. How much students learn from demonstrations depends on the way the demonstration is presented [32]. Several researchers have criticized the pedagogy of the traditional physics demonstration during which the instructor typically stands in front of the
Chapter 2: The Role of Physics Lecture Demonstrations in Conceptual Learning

lecture hall, actively manipulating the equipment and explaining the principles while the students passively watch and listen \[29,32\]. Studies have shown that, students learn little, if anything from demonstrations presented in a traditional manner \[32\]. Research on student learning from demonstrations suggests that traditional demonstrations may not effectively help students grasp the underlying concepts or recognize and correct scientific misconceptions that they may have \[29,32,34\]. There are two objectives to this study. The first is to examine the role of demonstrations in conceptual learning. The second is to provide useful insights into increasing the effectiveness of lecture demonstrations.

2.1.1 The Effect of Misconceptions on Observation

Students entering science classrooms have various preconceptions of how the physical world works. Students may have some conceptions that are unstable and highly context dependent \[35,37\] while other conceptions may be robust and appear to be in conflict with the concepts to be learned \[38\]. Conceptual learning of physics is often very difficult, particularly in instances when students have strongly held beliefs. One theory for why students typically fail to learn from teacher demonstrations is that their naive belief systems do not allow them to actually see what the demonstration is supposed to show \[29,34,39,40\]. Students “perceive science demonstrations from a perspective that differs from that of teachers and scientists” \[30\] and their ability to relate the scientific understanding to the observed event is limited \[41\]. Students who have misconceptions concerning the principles will be unlikely to see what the demonstration is showing; their perception of the demonstration will be
different from that of an expert [29]. Misconceptions can fundamentally affect how students perceive and interpret what they see and hear [36]. Several studies show that students’ observations reported after the execution of a science demonstration were in consonance with the students’ predictions of the outcomes even if these predictions contradicted the actual outcomes [39–42]. Gattis and Park (1997) [43] found that student misconceptions may actually be reinforced by demonstrations, rather than being corrected. There is evidence that this reinforcement may occur not only due to students’ incorrect observations but also, as a result of students’ false memories of the demonstrations.

2.1.2 Misconceptions and Memory

Just as the ability to observe a demonstration correctly may be hindered by students’ misconceptions, the ability to recall a demonstration may also be effected by students’ conceptual understanding of the underlying principles. Kraus (1997) [34] showed that some students remember demonstration outcomes that did not actually occur, especially when the outcome is in keeping with their own misconceptions. When shown a demonstration that does not fit with their own conceptual understanding, some students may alter their memory of the event rather than restructuring their model to accommodate for the contradictory observation [44].

Current theory on memory may explain how this ‘false memory’ is possible. Visual thinking (retrieving images from memories) is often driven more strongly by the conceptual knowledge we use to organize our images than by the content of the images themselves [45]. Instead of being stored as a continuous visualization
like a video recording, memory is reconstructed at the moment of recall [46]. Gaps in memory are filled in with information that comes from an individual’s conceptual model rather than by facts [46]. Therefore, an individual with an incorrect conceptual model could conceivably reconstruct the memory so that it supports their model rather than remembering what really took place.

2.1.3 Implications to Instruction: Conceptual Change

Theory suggests and research has shown that students’ pre-instructional beliefs may hinder their ability to both correctly observe and accurately remember demonstrations. On the other hand, conceptual change theory implies that demonstrations can actually be used to help students identify and confront misconceptions that they might have and thus force a conceptual change. For students to learn from demonstrations, a conceptual change must occur; students must alter or replace their misconceptions regarding the principle that the demonstration is designed to illustrate. Conceptual change and learning take place when students are forced to actively confront and explore their existing beliefs [47]. Some studies have shown that demonstrations that are designed to specifically address common areas of student difficulty are more effective in eliciting conceptual change than demonstrations that do not target areas of student misconceptions [34,44]. Implicitly, these studies suggest that instead of hindering the learning potential of demonstrations, misconceptions can be exploited through demonstrations to promote conceptual change.

Regardless of the role that misconceptions play in helping or hindering student learning, the effectiveness of a demonstration depends, at least in part, on the de-
livery method [32]. Passive observation of a demonstration has limited potential in promoting conceptual change in students’ understanding of a principle. Not only could students’ misconceptions hinder their ability to observe and perceive what the demonstration is designed to convey but the lack of engagement means that there is no opportunity for students to participate in social discourse practice (describing events and talking theory) [29].

Many successful research-based strategies for teaching physics involve a sequence of activities designed to actively address specific student misconceptions by engaging in social interactions [48, 49].

2.1.4 Elicit, Confront, Resolve

Elicit, confront, resolve is an effective instructional approach that promotes conceptual learning [50]. This approach first exposes students to a situation where they are likely to make a mistake, if they hold a given misconception. It then elicits their ideas about the situation. Students’ misconceptions are then confronted by exposing them to the error and helping them realize inconsistencies in their thinking. Inconsistencies are resolved by working through the reasoning required to reconcile their existing understanding with the new concept. In the context of demonstrations, the ‘elicit, confront, resolve’ cycle is better known as ‘predict, observe, explain’ (POE) [51]. Students’ misconceptions are elicited through their prediction of the outcome of a demonstration they are likely to find surprising. These misconceptions are confronted when students observe the demonstration and resolved when students are asked to explain their observation. The documented effectiveness of Interactive
Lecture Demonstrations (ILDs) is based on the POE strategy.

2.1.5 Interactive Lecture Demonstrations

In light of the research that has shown that traditional demonstrations are ineffectual in promoting conceptual change, there have been several implementations of teaching methodologies that increase students’ level of interaction and engagement during the lecture demonstration experience. Sokoloff and Thornton (1997) developed a series of Interactive Lecture Demonstrations (ILD’s) whose implementation has produced impressive gains in students’ conceptual change of the underlying physics principles. The ILD’s consist of a sequence of micro-based laboratory demonstrations during which students are asked to make predictions, answer questions, and describe and discuss the results with peers. The order and content of these sequences are based on the results of research in physics learning and are designed to address known areas of common misconceptions.

2.1.6 Role of Predictions

Prediction-making plays an important role in student learning. Couch et al. (2004) designed a study that combined Peer Instruction and the elicit, confront, resolve strategy into a methodology used to present lecture demonstrations. During this study, three different methods of presenting physics demonstrations were examined: (1) observe, the traditional approach in which students watch the demonstration and then hear the instructor’s explanation; (2) predict, in which students record their predictions of the demonstration outcome, observe the demonstration and then hear
the instructor’s explanation; and (3) discuss, in which students record predictions, observe the demonstration, discuss it with fellow students, and then finally hear the instructor’s explanation. The discuss mode provides students with the opportunity to elicit their ideas through the prediction stage, confront those who make incorrect predictions with their errors through the demonstration itself and then to explicitly resolve their misunderstanding through discussion with peers. In this study, students’ conceptual change was measured by comparing performance on a pre and post-test (multiple choice) that asked students to predict the outcomes of physical situations identical to the demonstrations presented in class. Comparison of pre and post-test scores showed that increasing the interactivity of the demonstration improves students’ learning outcomes. While the discuss mode group displayed the highest degree of conceptual change, the performance of the predict mode group performed almost as well; both modes producing significantly better gains that the traditional mode of instruction. No significant differences were found between students who observed the demonstrations and students who had not seen them at all. Students who were first asked to predict the outcome of the demonstration were significantly more likely to correctly identify the outcome of the demonstration \[32\]. For demonstrations to be effective, students must first predict the outcome.

Some studies of demonstrations have examined the relationship between students’ predictions and what they observe during a demonstration. Students often see what they predicted, even when this observation is different from what the instructor is actually demonstrating \[39\] \[41\] \[42\]. Hence, predictions are necessary, but they may not be sufficient.
2.1.7 Focus of Study

We study the role of predictions on conceptual learning. We examine how students’ predictions affect what they report having seen during a demonstration. We also examine how student predictions affect what they recall as the outcome of the demonstration at the end of the semester.

We find that roughly one out of every five observations (18%) of a demonstration is inconsistent with the actual outcome. Furthermore, conceptual learning depends on students accurately observing the outcome of the demonstration, regardless of whether their initial prediction is correct or incorrect. Contrary to previous studies, students who make an incorrect prediction are as likely to accurately report the outcome of the demonstration as those who make a correct prediction. This may be due to the surprisingly high number of students who predict correctly but then fail to accurately report the outcome of the demonstration. Indeed though 20% of students who predict incorrectly also report the incorrect outcome, 16% of students who predict correctly also report the outcome incorrectly. We identify two situations that are associated with successfully observing and remembering a demonstration: 1) first predicting the outcome before seeing the demonstration, (regardless of whether the prediction is correct or not) and 2) having a basic prior understanding of the concepts underlying the demonstration.
Chapter 2: The Role of Physics Lecture Demonstrations in Conceptual Learning

2.2 Methods

Students in this study were registered in one of two introductory physics courses for non-majors at two large research institutions: a mechanics class for engineering students (N = 201) and an electricity and magnetism class for engineering and pre-medical students (N = 91). Lecture demonstration data were collected from 22 different demonstrations in these two classes (10 from the mechanics class and 12 from the E&M class). Throughout the semester, the instructor presented demonstrations during the lectures as a regular part of the course.

While the analysis presented represents results pooled from both classes, each class was first analyzed separately. We combined the results after establishing identical trends in each. Students in both classes were told to follow the same procedure when it came to each demonstration. Table 2.1 presents a summary of this procedure. Some students chose not to predict the outcome of the demonstration. In their response to the outcome question, these students reported being present and having seen the demonstration in class. We compare these to students who did make a prediction.

We examined students’ predictions to the outcomes of lecture demonstrations and what they report as having observed during the demonstration both directly afterwards and, at the end of the semester.

Immediately before and immediately after each demonstration, students were required to answer 2-4 multiple-choice conceptual questions. Pre- and post-demonstration questions were different, but were designed to be of similar difficulty. Both sets of questions were designed to probe students’ understanding of the physics underlying
Chapter 2: The Role of Physics Lecture Demonstrations in Conceptual Learning

Table 2.1: Summary of data collection procedure.

<table>
<thead>
<tr>
<th>Measurement Instrument</th>
<th>Timing</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-class</td>
<td></td>
</tr>
<tr>
<td>1) 2-4 conceptual Qs</td>
<td>before demo</td>
</tr>
<tr>
<td>2) predict Q</td>
<td></td>
</tr>
<tr>
<td>3) 2-4 conceptual Qs</td>
<td>after demo</td>
</tr>
<tr>
<td>On-line</td>
<td></td>
</tr>
<tr>
<td>4) Q on demo outcome</td>
<td>1 day after demo</td>
</tr>
<tr>
<td>5) Q on demo outcome</td>
<td>end of semester</td>
</tr>
</tbody>
</table>

each demonstration. To measure conceptual learning, we compare students’ performance on the pre-demonstration questions with their performance on the post-demonstration questions. Appendix 1 provides examples of the types of pre- and post-conceptual questions that were used to probe students’ understanding of the physics underlying the demonstration. Also provided are examples of the demonstrations used in this study. Responses to both sets of conceptual questions were recorded via a web-based polling system during class.

Students were asked to predict the outcome of a demonstration by responding to a multiple-choice question with choices that spanned the full range of possible outcomes. Students were given a few minutes to think and record their predictions, and were told not to discuss with their classmates. Predictions were recorded with the same web-based system in class.

Students were asked what they had observed during the demonstration twice after each demonstration. Right after lecture students had a day to answer online (1) a multiple-choice question asking what they observed as the outcome of the demon-

20
stration and (2) a free-response question asking them to explain their understanding of the physics behind the demonstration. Students who missed the demonstration in-class could select I did not see this demonstration one of the response options for the multiple-choice question. The same two questions were asked again at the end of the semester. We compare what students report having seen after the demonstration to what they remember at the end of the semester.

2.3 Results and Discussion

Figure 2.1 shows aggregate results of students’ responses to the outcome questions asked after class and at the end of the semester. The average percentage of correct outcomes is shown for students who (1) failed to make a prediction, (2) predicted correctly and, (3) predicted incorrectly. We confirmed that the population in each group was equivalent in aptitude by calculating the average final grade for each of the no prediction, predict incorrectly and predict correctly groups, which were 77.6%, 78.4%, and 78.8%, respectively. The difference in final grade between the no predict and other two (predict) groups is not significant at the $p < 0.05$ level. Using data from the web-based polling system, we determined that students who did not make predictions had answered other questions during the same classes as the demonstrations. In the outcome question after the demonstration, these students did not select the I did not see this demonstration option in the outcome question. Hence, students who did not make a prediction were present during the class and claimed to have seen the demonstration. It is possible that these students did not predict the outcome of the demonstration because they were inattentive when the prediction
question was posed, did not know the answer or simply failed to register their answer in time.

Figure 2.1: Average percentage of correct response to the outcome question asked 1 day after the demonstration (light grey) and at the end of the semester (dark grey), for students who: made no prediction (n), predicted the demonstration correctly (c) or incorrectly (i).

Figure 2.1 validates previous research that pointed to the importance of having students first predict the outcome of the demonstration [32]. A comparison of the no predict group to the other two (predict) groups indicates that students who predict are significantly more likely to correctly report the outcome of a demonstration, regardless of whether asked within a day of seeing the demonstration or several weeks
later. Making a prediction, regardless of whether it is correct or not, appears to play an important role in correctly observing the event and a less important role in remembering it.

Especially noteworthy is that students who predict incorrectly answer the outcome question correctly at approximately the same rate as students who predict correctly. This finding goes against aforementioned studies [39,41,42] that have suggested students’ observations are often in consonance with their incorrect predictions. Surprisingly, a high percentage (16% on average) of students who do predict correctly, still report the outcome of the demonstration incorrectly when asked within one day of the event.

Also of note is the finding that there is no significant difference in the percentage of correct outcome response between the two instances in time (within a day of having seen the demonstration compared to at the end of the semester). Assuming they make a prediction, students who state the outcome correctly in the first place are likely to remember it when asked later in the semester, regardless of whether their prediction was correct or incorrect.

In addition to the role of prediction-making, we also looked at students’ level of conceptual understanding. Students were grouped according to their performance on the pre-demonstration conceptual questions. Figure 2.2 shows how students at different levels of pre-demonstration conceptual understanding respond to the outcome question within a day of the demonstration and at the end of the semester. Figure 2.2 compares students at three different levels of performance on these questions: (1) bottom third of the class (2) middle third of the class (3) top third of the class. Stu-
Figure 2.2: Average percentage of correct response to the outcome question asked 1 day after the demonstration (light grey) and at the end of the semester (dark grey), for students at three different levels of performance on the pre-demonstration conceptual questions; bottom third of the class (B), middle third of the class (M), and top third of the class (T).

Students with a strong understanding of the concepts before seeing the demonstration correctly report the outcome at the end of the semester at a significantly higher rate than students who have a weak conceptual understanding. Students whose pre-demo conceptual performance is in the top third of the class, correctly report the outcome of the demonstration 21% ($p < 0.001$) more frequently than students in the bottom third of the class.
To measure the role that demonstrations play in conceptual learning, we compare students’ performance on the pre and post-demonstration conceptual questions. Figure 2.3 compares the (normalized) post-demonstration conceptual performance of students who responded correctly to the initial outcome question to students who responded incorrectly. To control for pre-demonstration conceptual test performance, we make this comparison within the three different levels of performance on the pre-demonstration questions.
Table 2.2: Standardized coefficients for linear regression model predicting performance on post-demonstration conceptual questions based on correct/incorrect observation of demonstrations. Model controls for students’ pre-demonstration conceptual performance.

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct/incorrect demo outcome</td>
<td>0.33**</td>
<td>**&lt;0.0001</td>
</tr>
<tr>
<td>pre-demo conceptual performance</td>
<td>0.27**</td>
<td>*&lt;0.005</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.93</td>
<td></td>
</tr>
</tbody>
</table>

**$p <0.0001$ * $p <0.005$**

Table 2.2 shows the standardized regression coefficients for the linear regression model used to predict performance on post-demonstration conceptual questions based on whether students’ correctly stated the outcome of the demonstration. Student’s pre-demonstration conceptual performance is used as a control in this model.

Students who correctly state the outcome of the demonstration experience significantly higher conceptual learning ($p<0.005$) than students who do not state the outcome correctly. Within the top third in pre-demonstration conceptual understanding, students who correctly identify the outcome of the demonstration score 17% higher on the post-demonstration conceptual test than students who incorrectly identify the outcome of the demonstration.

### 2.4 Conclusion

Our results replicate earlier findings on the importance of having students predict the outcome of a demo \[32\] and lead to three new conclusions. The first is that when asked within a day of seeing a demonstration, students report an incorrect out-
come 18% of the time, regardless of whether they predict the outcome correctly or incorrectly. The second is that for a demonstration to lead to conceptual learning, it is important that the student observes it correctly in the first place. In addition, students are more likely to observe the demonstration correctly if they a) make a prediction first and b) have some conceptual understanding of the underlying physics beforehand. These findings support the importance of having students predict the outcome of a demonstration, regardless of whether they predict correctly or not. Furthermore, demonstrations are most effective in promoting learning when students have at least a basic level of conceptual understanding beforehand. In light of these findings, we recommend that demonstrations be integrated into a learning sequence that helps students develop background knowledge first. These findings support the effectiveness of POE strategies such as Interactive Learning Demonstrations [48], which emphasize both prediction making and conceptual scaffolding of the demonstration.
Chapter 3

Response Switching and Self-Efficacy in Peer Instruction Classrooms

Peer Instruction, a well-known student-centered teaching method, engages students during class through frequent questioning and is often facilitated by classroom response systems (CRSs). The central feature of any Peer Instruction class is a conceptual question designed to help resolve student misconceptions about subject matter. We provide students two opportunities to answer each question once after a moment of individual reflection and then again after discussion with a peer. The second round provides students the choice to “switch” their original response to a different answer. The percentage of right answers typically increases after peer discussion: most students who answer incorrectly in the individual round switch to the correct answer after the peer discussion. However, for any given question there are also students who switch their initially right answer to a wrong answer and students who switch their initially wrong answer to a different wrong answer. In this study, we analyze
switching responses over one semester of an introductory electricity and magnetism course taught using Peer Instruction at Harvard University. Two key features emerge from our analysis: First, switching correlates with academic self-efficacy. Students with low self-efficacy switch their answers more than students with high self-efficacy. Second, switching also correlates with the difficulty of the question; students switch to incorrect responses more often when the question is difficult. These findings indicate that instructors may need to provide greater support for difficult questions, such as providing cues during lectures, increasing times for discussions, or ensuring effective pairing (such as having a student with one right answer in the pair). Additionally, the connection between switching and self-efficacy motivates interventions to increase student self-efficacy at the beginning of the semester, in the hope that this will improve student learning in Peer Instruction classrooms.

3.1 Introduction

Peer Instruction, a student-centered teaching methodology developed in the 1990s, engages students during class through a sequence of questioning and discussion. Questions are generally conceptual in nature and probe students’ ability to apply their understanding to solve conceptual problems. These questions, called “ConcepTests”, are the centerpiece of Peer Instruction. Students first respond to a ConcepTest individually (first round) and then respond again to the same question after discussing with a peer (second round). The students’ responses are typically recorded via classroom response systems, which allow the instructor to track the class-wide percentage of right answers between the two rounds of questioning. This percentage almost al-
Chapter 3: Response Switching and Self-Efficacy in Peer Instruction Classrooms

ways increases for a given question after peer discussion (that is, between the first and second rounds of questioning). However, for any given question, there are also both students who switch their initially right answer to a wrong answer and students who switch their initially wrong answer to a different wrong answer. Understanding the difference between the different types switching helps provide insight into student cognition in Peer Instruction environments. We pose three research questions. First, how often does switching occur and in which direction? Second, what is the relationship between ConcepTest switching and pre-determined student characteristics, specifically gender, pre-course physics knowledge, and pre-course self-efficacy? Third, what is the relationship between ConcepTest difficulty and student switching? While the relationship between pre-course knowledge and student achievement is well studied, there is increasing interest in the relationship between non-cognitive dimensions, such as self-efficacy and academic performance. Self-efficacy refers to an individual’s belief that s/he can successfully complete a task \(^2\). Self-efficacy is a strong predictor for performance in science courses \(^3\,^6\) as well as science career choices \(^4\,\,5\,7\,8\). Self-efficacy also influences a number of factors that are relevant to learning in a Peer Instruction environment, such as perseverance and self-regulated learning \(^9\). Students with higher self-efficacy are more persistent, harder working, participate more readily, and experience fewer negative emotions in the face of difficulty than students with lower self-efficacy \(^53\). In their study on the effects of self-efficacy on student behavior during conceptual learning, Bouffard-Bouchard et al. (1991) found that “efficacious students were better at monitoring their working time, more persistent, less likely to reject correct hypotheses prematurely, and better
at solving conceptual problems than inefficacious students of equal ability” (Zimmer-
man, 2000, pg 87). In this paper, we show that ConcepTest switching is related to
students’ pre-course self-efficacy. Given this relationship between student self-efficacy
and switching, it is possible that improving a student’s self-efficacy would enhance
the effectiveness of Peer Instruction for that student. Recent work has identified a
technique for identifying events in small group settings that impact self-efficacy in
physics learning [54]. This work has exciting implications for improving students’
self-efficacy in Peer Instruction environments.

3.2 Methods

We gathered ConcepTest (CT) response data over an entire semester in an intro-
ductive, calculus-based electricity and magnetism class taught using Peer Instruction
at Harvard University, by a professor with twenty years of experience teaching with
this method. CTs are short conceptual questions that focus on a single topic [49]. The
class had 91 students (50 males and 41 females), the majority of whom were engineers
or premedical students. The class met twice a week for 90 minutes and during each
class, somewhere between five and 15 CTs were posed. In total, 83 CTs were posed
over the course of the entire semester. Students answered each CT in two rounds of
questioning by entering their responses via Learning Catalytics, an online classroom
response system. Students did not receive credit based on the correctness of their an-
swer; rather, credit was only given based on whether the student responded at all. We
divided each pair of CT responses into one of five transition categories according to
the response pattern during the two rounds of Peer Instruction; (1) right-right (RR),
Chapter 3: Response Switching and Self-Efficacy in Peer Instruction Classrooms

the question is answered correctly during both rounds, (2) wrong-to-right (WR), the question is answered incorrectly in round 1 and correctly in round 2, (3) right-to-wrong (RW), the question is answered correctly in round 1 and incorrectly in round 2 (4) wrong-wrong same (WW-S), the question is answered with the same incorrect response in both rounds, and (5) wrong-wrong different (WW-D), the question is answered with a different wrong response in both rounds. Only questions answered both rounds were included in the analysis. If a student responded to a question in round 1 but not in round 2 (or vice versa), we dropped that data point from our analysis. We applied a two-parameter Item Response Model \cite{55} to the first round of responses across all items to estimate the difficulty of each item (b-parameter). The Item Response Model helps to scale the difficulty of each question to support the generalization of the item difficulties to other populations of students. The Conceptual Survey of Electricity and Magnetism (CSEM) \cite{10} and a self-efficacy survey were both administered twice, once as pre-tests after the second class in the semester, and again as post-tests at the end of the semester (after the final exam but before students received their final grade). The self-efficacy survey used in this study was developed by two of the authors and is based on the Sources of Self-efficacy in Science Courses (SOSESC) survey \cite{56}. This survey is comprised of 21 items and a five-point Likert scale. Students are asked if they “strongly agree”, “agree”, are “neutral”, “disagree”, or “strongly disagree” with statements about whether they think they will be successful in a number of physics-related tasks (e.g., solving difficulty physics problems or communicating physics successfully to a peer). The complete survey can be found in Appendix 2.
Chapter 3: Response Switching and Self-Efficacy in Peer Instruction Classrooms

The Pearson point biserial coefficient measures the consistency between an item and the test as a whole. An item with a low point biserial indicates that the item is measuring something that is inconsistent with the overall survey. A survey designed to measure a single construct (such as self-efficacy) should have items with point biserial coefficients $>0.2$ [57]. All items on our self-efficacy survey have point biserial coefficients greater than 0.2 except for item 13, which was dropped from analysis. Cronbach Alpha, a measure of overall reliability, is the overall correlation between survey items [58]. An instrument designed to measure a single construct is considered reliable when it has a Cronbach Alpha $>0.7$ [58]. Our self-efficacy survey has a Cronbach Alpha of 0.88 for the pre-test and 0.85 for the post-test.

3.3 Results

3.3.1 Descriptive Statistics on Switching

Figure 3.1 shows the extent to which students switch their CT responses between the first and second round of questioning. Sub-plot A shows the fraction of CTs that are switched (in any direction) out of the total number of CTs that each student answers. Figures 3.1 B-D show the fraction of CTs that are switched in a specific direction out of the total number of CTs that are switched (B: wrong to right; C: wrong to different wrong; D: right to wrong). On average, students switch 44% of the CTs they answer over the course of the semester. Of this, 73% is switching from wrong to right, 17% is switching from wrong to a different wrong and 10% is switching from right to wrong.
Figure 3.1: A: Fraction of all CT responses that are switched (top left), B: fraction of all switched responses that are switched from wrong to right (top right), C: fraction of all switched responses that are switched from wrong to a different wrong (bottom left), and D: fraction of all switched responses that are switched from right to a wrong (bottom right).

3.3.2 Normalizing Switching

When we measure switching as we did in figure 3.1 (as a fraction of switches out of all the questions responded to), the switching measure is confounded with the frequency of right (or wrong) answers in round 1. Normalizing the variables with respect to the response in round 1 provides us with an adjusted measure of the switch independent of how many times a student was right (or wrong) in round 1. Figure
Figure 3.2 illustrates the need for normalization. In Figure 3.2, each point represents the relationship between the number of items a student answered incorrectly in round 1 and the fraction of answered items that were switched from wrong to right. When the WR transition is not normalized with respect to the round 1 response, the number of wrong, round 1 responses confounds the number of wrong to right switches. To illustrate this point, consider the two students highlighted in Figure 1, both of whom switch approximately 50% of their responses from wrong to right. Student 1 answered less than 20 items incorrectly to begin with and therefore switched to the right answer fewer than 10 times. Student 2, on the other hand, had more than forty wrong responses in round 1 and therefore switched from wrong to right more than 20 times. Normalizing adjusts the response transition (WR) to account for the frequency of initial wrong answers.

To calculate each of the normalized transition variables we first summed the number of times each student’s response fell into each of the five transition categories. Then, we normalized each of these sums to express them as a percentage of the times that the first response was wrong or right. The normalized version of the RR transition was computed by dividing the total number of questions a student was right in both rounds by the number of questions they had the right answer in round 1. The WR transition was computed by dividing the total number of questions a student switched their answer from wrong to right by the number of questions they had the wrong answer in round 1. The RW transition was computed by dividing the total number of questions a student switched their answer from right to wrong by the number of questions they had the right answer in round 1. A summary of how these
Figure 3.2: Fraction of all CT responses that each student switches from wrong to right plotted as a function of the number of questions the student answers incorrectly in round 1. Students 1 and 2 both switch approximately 50% of the questions they answer from wrong to right. These two students are not directly comparable, however, because student 1 has far fewer incorrect answers in round 1 than student 2. Normalizing with respect to the number of incorrect answers in round 1 allows us to compare these two students. Normalized transition variables are calculated is provided in Appendix 1.

3.3.3 Switching and Self-efficacy

We find that CT switching behavior is a function of students’ pre-course self-efficacy. Students with low self-efficacy are both more likely to switch and more likely
Chapter 3: Response Switching and Self-Efficacy in Peer Instruction Classrooms

to switch in a ‘negative’ direction (from right to wrong and from wrong to a different wrong) than students with high self-efficacy. Students with high self-efficacy are much more likely to switch from wrong to right than students with low self-efficacy. Figure 3.3 shows the average normalized percentage of CTs switched for students with low and high self-efficacy.

![Figure 3.3: Average percent of questions switched in each of the three directions (wrong to right, right to wrong, wrong to different wrong) normalized with respect to the first response, for students with high and low self-efficacy](image)

Figure 3.3: Average percent of questions switched in each of the three directions (wrong to right, right to wrong, wrong to different wrong) normalized with respect to the first response, for students with high and low self-efficacy

Table 3.1 displays the standardized regression parameters and significance metrics for two models, each of which predicts the proportion of CT questions switched from right to wrong (RW), wrong to right (WR) and wrong to a different wrong
Table 3.1: Standardized coefficients for linear regression models predicting proportion of CT questions switched from right to wrong (RW), wrong to right (WR) and wrong to a different wrong (WW-D), all normalized with respect to the first response. Model 1 controls for pre-course student self-efficacy only whereas Model 2 controls for both pre-course student self-efficacy and CSEM-scores.

(WW-D), normalized with respect to the first response. In each set of two models, Model 1 controls for pre-course student self-efficacy only whereas Model 2 controls for both pre-course student self-efficacy and CSEM-scores. Students with high pre-course self-efficacy switch from right to wrong and wrong to a different wrong significantly less often, and switch from wrong to right significantly more often than students with low self-efficacy (\( p<0.001 \)). This is true even when incoming physics knowledge is controlled for, indicating that the self-efficacy measurement is not simply a proxy for incoming physics knowledge. The magnitude of the standardized coefficients represents their relative predictive power in each of the models; a comparison of the self-efficacy coefficient to the CSEM coefficient in Model 2 indicates that self-efficacy is more predictive of switching than incoming physics knowledge.
Figure 3.3 shows the predicted proportion of questions switched in each of the three directions, based on model 2, for students with high and low self-efficacy. Students with self-efficacy scores at least one standard deviation greater than the mean were classified as those with high self-efficacy and compared to students with low self-efficacy (scores less than one standard deviation less than the mean). Students with high self-efficacy switch their answers from wrong to right more than students with low self-efficacy ($p<0.05$). Students with low self-efficacy switch their answers from right to wrong ($p<0.005$) and from wrong to a different wrong ($p<0.05$) more than students with high self-efficacy.

Students’ responses to two individual items on the self-efficacy survey correlate strongly with switching the right answer to a wrong answer. Figure 3.4 shows average right to wrong switching (normalized) for students with different levels of agreement/disagreement with the statements “I usually don’t worry about my ability to solve physics problems” (item 10) and “I can communicate science effectively” (item 20). A one-way analysis of variance indicates that students who strongly disagree with item 10, switch from right to wrong significantly more than students who agree (or disagree less strongly) with that item ($p<0.001$). Similarly, students who disagree with the item 20, switch from right to wrong significantly more than students who agree with that item ($p<0.05$). These relationships are significant even after controlling for students’ incoming CSEM scores. Therefore, independent of their actual physics ability, students with a low assessment of their problem solving and communicating science abilities are significantly more likely to switch their answers from right to wrong than students with a high assessment of those abilities.
Chapter 3: Response Switching and Self-Efficacy in Peer Instruction Classrooms

Figure 3.4: Average right to wrong switching (normalized) for students with different levels of agreement with the statements "I can communicate science effectively" (left plot) and "I usually don’t worry about my ability to solve physics problems" (right plot).
3.3.4 Switching and Gender

Table 3.2 shows that a gender difference exists in the fraction of CTs that are switched. Female students switch 15% more overall (p<0.05) and 45% more from right to wrong (p<0.05) than male students — but this gender difference disappears when controlling for self-efficacy. The reason this difference disappears is female students have lower pre-course self-efficacy than male students. Female students score 10% lower than male students on the pre-course self-efficacy survey. We also find that females switch less from wrong to right and more from wrong to a different wrong than males, although neither of these differences is statistically significant. Table 3.2 displays the standardized coefficients for three different linear regression model predicting the ratio of overall switching, right to wrong switching (normalized), wrong to right switching (normalized) and wrong to different wrong switching (normalized). Model 1 controls for gender only, Model 2 controls for gender and pre-course self-efficacy, and Model 3 controls for gender, self-efficacy and pre-course CSEM scores. Gender alone is a significant predictor for overall switching and right to wrong switching (p<0.05) but when self-efficacy is added to the model, it is no longer a significant predictor in either case.

3.3.5 Switching and CT Difficulty

To investigate the relationship between ConceptTest difficulty and student switching, we shift our focus to item level switching by looking at what percentage of students switch for each individual ConceptTest. Figure 3.5 shows the percent of students who switch (in any direction) as a function of CT difficulty. Each point on Figure 3.5
Table 3.2: Standardized coefficients for linear regression models predicting proportion of CT questions: switched out of all the questions answered (switched), switched from right to wrong out of all the questions answered correctly in the first round (RW normalized), switched from wrong to right out of all the questions answered incorrectly in the first round (WR normalized), and switched from wrong to a different wrong out of all the questions answered incorrectly in the first round (WW-D normalized). Model 1 controls for gender, Model 2 controls for gender and pre-course student self-efficacy, and Model 3 controls for gender and both pre-course student self-efficacy and CSEM-scores.

<table>
<thead>
<tr>
<th></th>
<th>Switched</th>
<th>RW (normalized)</th>
<th>WR (normalized)</th>
<th>WW-D (normalized)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
<td>Model 3</td>
<td>Model 1</td>
</tr>
<tr>
<td>Gender</td>
<td>0.47*</td>
<td>0.29</td>
<td>0.21</td>
<td>0.43*</td>
</tr>
<tr>
<td>Self efficacy</td>
<td>-0.29**</td>
<td>-0.18</td>
<td>-0.28</td>
<td>-0.28*</td>
</tr>
<tr>
<td>CSEM</td>
<td>-0.46***</td>
<td>-0.19</td>
<td>-0.13</td>
<td>-0.13</td>
</tr>
<tr>
<td>R²</td>
<td>0.05</td>
<td>0.13</td>
<td>0.40</td>
<td>0.04</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.98</td>
<td>0.95</td>
<td>0.76</td>
<td>0.08</td>
</tr>
</tbody>
</table>

*p < 0.05  ** p < 0.01  *** p < 0.001
Figure 3.5: Fraction of students who switch (in any of the three directions) plotted for each individual ConcepTest as a function of CT difficulty (as estimated by a 2PL Item Response Theory model).

represents an individual CT.

The difficulty of each CT \((b\)-parameter\) was estimated using a 2PL Item Response Theory model [55] and this was plotted against the fraction of students that changed their response from the first to the second round. As the difficulty of the item increases, so too does the percentage of students who switch their response. With increasing item difficulty, students are more likely to switch from right to wrong \((p<0.001)\) or from wrong to a different wrong \((p<0.05)\) and less likely to switch from wrong to right \((p<0.05)\). The correlation between switching and CT difficulty is 0.54
Table 3.3: Correlation between the difficulty of each item (as estimated using a 2PL IRT model) and the fraction of students who, for that item, have a different wrong answer in each round (WW-D), switch from the right answer to a wrong answer (RW), switch from a wrong answer to the right answer (WR), have the same wrong answer in both rounds, and have the right answer in both rounds.

<table>
<thead>
<tr>
<th></th>
<th>RW</th>
<th>WR</th>
<th>WW-D</th>
<th>WW-S</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>item difficulty</td>
<td>0.38***</td>
<td>-0.24*</td>
<td>0.25*</td>
<td>0.10</td>
<td>-0.42***</td>
</tr>
</tbody>
</table>

(p<0.001). Table 3.3 shows the correlations between CT difficulty (b-parameter) and the fraction of students who, for each question: (1) switch from right to wrong divided by the number of students who answered correctly in round 1, (2) switch from wrong to right divided by the number of students who answered incorrectly in round 1, (3) switch from wrong to a different wrong divided by the number of students who answered incorrectly in round 1, (4) have the same wrong answer in both rounds divided by the number of students who answered incorrectly in round 1, and (5) have the right answer in both rounds divided by the number of students who answered correctly in round 1. Table 3.3 shows that switching is related to the difficulty of the item. As the difficulty of the item increases, students are more likely to switch from right to wrong (p< 0.001) or from wrong to a different wrong (p<0.05) and less likely to switch from wrong to right (p<0.05).
3.4 Discussion

3.4.1 Class-wide Switching

The class-wide percentage of correct answers in each round of questioning provides real-time feedback of student understanding and is used, by the instructor, to guide the class during Peer Instruction. Ideally, students who respond incorrectly initially will switch to the right answer after having a discussion with their peers. Similarly, in the ideal situation students who initially respond correctly will not switch to a wrong answer after the discussion. We find that, on average, students switch on 44% of the CTs they answer and that the vast majority of these switches (73%) is from wrong to right. Although this percentage is high, there is still a significant proportion of CT switching (27%) in directions that are negatively associated with student learning (right to wrong and wrong to different wrong).

3.4.2 Switching and self-efficacy

Students with low self-efficacy have been shown to be less persistent and experience more negative emotions in the face of difficulty than students with high student efficacy \cite{53}. Bouchard et al. (1991) found that students with low self-efficacy were more likely to reject correct hypotheses prematurely and struggle with conceptual problem solving than students of the same ability with high self-efficacy. We find that students who have low confidence in their ability to solve problems and communicate science score are also significantly more likely to switch the right answer to a wrong one. The data shows that students with a low assessment of their ability to
communicate and solve problems frequently switch from right to wrong. This behavior is, at best, frustrating for students, and, at worst, contributing to an even further decrease in academic self-efficacy. Recent work has shown student's interactions in small group settings can impact self-efficacy in physics learning [54]. Interventions to increase student self-efficacy at the beginning of the semester could improve student learning in interactive environments such as Peer Instruction classrooms.

### 3.4.3 Switching and CT difficulty

It is important for instructors to understand that they have some measure of control over the switching that occurs in their classrooms via the difficulty of the ConceptTests that they select. Understanding that students switch to the wrong response more often with difficult questions is informative because it indicates that instructors may need to provide better scaffolding for those questions. Instructors should try to support students through more difficult questions by providing mini-lectures, more pre-class reading materials, or by keeping the polls open longer. Instructors should also consider scaffolding the most difficult CTs by building up to them with a series of less difficult questions.

### 3.5 Conclusion

Two results from our analysis of student ConceptTest switching behavior in a Peer Instruction environment have implications for classroom practice. The first is that CT switching behavior is a function of students’ pre-course self-efficacy. Students with low pre-course self-efficacy are more likely to switch to a wrong response and
less likely to switch to a right answer. Second, as the difficulty of the item increases, students are more likely to switch to a wrong response and less likely to switch to the right response. The strong connection between CT switching and self-efficacy has important implications for Peer Instruction. Interventions to increase student self-efficacy at the beginning of the semester might improve students’ experiences during Peer Instruction and help students take better advantage of this teaching strategy.
Chapter 4

Conceptual question response times
in Peer Instruction classrooms

4.1 Introduction

Classroom Response Systems (CRSs) are widely used in interactive teaching environments as a way to engage students by asking them questions. Previous research on the time taken by students to respond to conceptual questions has yielded insights on how students think and change conceptions. We measure the amount of time students take to respond to in-class, conceptual questions (ConcepTests or CTs) in two introductory physics courses taught using Peer Instruction and use Item Response Theory to determine the difficulty of the CTs. We examine response time differences between correct and incorrect answers both before and after the peer discussion for CTs of varying difficulty. We also determine the relationship between response time and students’ performance on a standardized test of incoming physics knowledge,
pre-course self-efficacy, and gender. Our data reveal three results of interest: First, response time for correct answers is significantly faster than for incorrect answers, both before and after peer discussion, especially for easy CTs. Second, students with greater incoming physics knowledge and higher self-efficacy respond faster in both rounds. Third, there is no gender difference in response rate after controlling for incoming physics knowledge scores, although males register significantly more attempts before committing to a final answer than do female students. These results provide insight into effective CT pacing during Peer Instruction. In particular, in order to maintain a pace that keeps everyone engaged students should not be given too much time to respond. Once around 80% of the answers are in, the ratio of correct to incorrect responses rapidly approach levels indicating random guessing and instructors should close the poll.

When used effectively, Classroom Response Systems (CRSs) can facilitate student engagement in science classrooms in ways that would not be possible otherwise [59, 60]. From low-tech versions, such as flashcards or white boards, to higher-tech, web-based response systems, such as Poll Everywhere or Learning Catalytics, the popularity of CRSs has also opened up an active area for educational researchers to pursue questions about their use and student learning outcomes. Peer Instruction is one popular research-based instructional strategy that leverages the power of CRSs to promote student learning [49]. In Peer Instruction, students use CRSs to respond to ConcepTests (CTs) both before and after discussing their answers with their peers for 2 to 5 minutes [15]. ConcepTests are short conceptual questions that focus on a single topic [49]. During Peer Instruction, students first respond individually to a
Chapter 4: Conceptual question response times in Peer Instruction classrooms

CT and then respond a second time to the same CT after discussing their responses with a peer. Instructors can use CT response data to inform teaching decisions in real time. Researchers suggest that Peer instruction is a superior teaching strategy in promoting student conceptual understanding and problem solving skills compared to the traditional lecture [15, 49, 61–64]. The purpose of this research is to gain insight into student thinking in Peer Instruction environments using one specific unit of analysis related to CTs: response time.

Despite the fact that Peer Instruction is a research-based pedagogy shown to be highly effective in promoting student learning, there is room for improvement particularly in implementation. Response times can provide insight into student thinking during Peer Instruction. Previous research has shown that when students answer conceptual questions with misconception-like responses, they tend to respond more quickly than those answering correctly [65]. With this insight, instructors can improve the implementation of Peer Instruction. For example, an examination of response time could inform CT pacing and help determine the optimal amount of time instructors should keep questions open for student response. Providing instructors with guidelines for appropriate CT pacing allows them to implement Peer Instruction more efficiently by optimizing class time.

We pose two research questions. First, what are the response time differences between correct and incorrect answers, before and after peer discussion for CTs of varying difficulty? Second, what is the relationship between response time and students’ performance on a standardized test of incoming physics knowledge, pre-course self-efficacy, and gender?
Physics education researchers have previously conducted studies of response times to conceptual questions to gain insight into student learning. Using standardized conceptual questions delivered as a pre-post course conceptual survey, Lasry et al. (2013) showed that response times for incorrect answers are longer than for correct responses [66]. Our study differs from Lasry et al. (2013) in that we examine in-class formative conceptual questions versus pre-post course surveys. Richardson et al. (2013) examined the relationship between gender and CT response time in a Peer Instruction classroom and found no statistically significant difference in response time between males and females [67]. However, Richardson et al. (2013) also found that males were more likely than females to change their answer within a single round before committing to a final response [67]. In this study of CT response times, we extend this work by combining response times with student learning data to gain further insight into student thinking in Peer Instruction environments.

4.2 Methods

We collected student responses and response times for ConcepTests from one semester in two introductory electricity and magnetism classes at Queen’s University (Kingston, Ontario; N = 48) and Harvard University (Cambridge, Massachusetts; N = 93). Both classes used Peer Instruction with between five and 15 CTs per class. Over the course of the semester, 101 CTs were given at Queen’s and 74 at Harvard. In both classes, students did not receive credit for getting the answer right. Credit was only given for responding to the questions. Students answered each CT in two rounds of questioning by entering their responses via a classroom response
system (CRS) (iClicker at Queen’s University and Learning Catalytics at Harvard). The CRS recorded a timestamp for each student’s final response. Response times were computed as the difference between the time the question was delivered to the students (via the CRS) and the time of each student’s final response. Regardless of the number of responses a student entered for a single question, only the time to the final response was recorded. For each question, we analyzed the time taken to respond before and after the discussion. The response time before the discussion is the amount of time it takes students to respond individually to the question the first time they see it. After students discuss answers with each other, the instructor presents the question again and opens the polling for a second time. The response time after the discussion is the amount of time that students take to respond after the polling is opened the second time. We did not constrain how long each question would stay open because some questions take longer to parse than others. We also chose to give each instructor the control of how much time each question stayed open. Students often continue their discussion past the time when the poll is reopened for the second round. Consequently, some questions stayed open much longer than others. In some rare cases, students were given more than five minutes to respond to a single question, especially in the second round. To control for the variability in response times across questions, we focus our analysis on the difference in the time taken to respond correctly and incorrectly. The iClicker system used at Queen’s also records the number of times a student responds to a CT before registering their final answer and, for this subset of the data, we included this in our analysis.

The Conceptual Survey of Electricity and Magnetism (CSEM) [10] and a self-
efficacy survey were both administered twice in each class, once as pre-tests at the beginning of the semester and again as post-tests at the end of the semester. Self-efficacy is a person’s belief that s/he can be successful when performing a situation-specific task [2], and students’ self-efficacy has been shown to be a strong predictor for perseverance and success in science [56]. The self-efficacy survey used in this study was developed at Harvard, and based on the Sources of Self-efficacy in Science Courses (SOSESC) survey [56]. This survey asked students to rank, on a five-point scale, the extent to which they believe they will be successful in a number of physics-related tasks (e.g., how successful will they be solving difficult physics problems or communicating physics successfully to a peer).

A two-parameter Item Response Model [55] was applied to the first round of responses across all items to determine the difficulty of each item (b-parameter). Items were grouped into two categories according to difficulty. Items for which b < 0 were classified as easy and items with b \geq 0 were classified as hard. In the Queen’s dataset, 59 out of the 101 CTs were classified as easy and the remaining 42 as hard. In the Harvard dataset, 38 of the 74 CTs were classified easy and the remaining 36 as hard.

4.3 Results

Table 4.1 shows the average time taken for correct and incorrect answers on ConcepTests both before \langle t_{before} \rangle and after \langle t_{after} \rangle peer discussion at Queen’s and Harvard University (QU and HU respectively). While the questions used at each institution were ConcepTests on electricity and magnetism, the same questions were
Table 4.1: Average time taken for correct and incorrect answers on ConceptTests, before and after Peer Instruction.

<table>
<thead>
<tr>
<th></th>
<th>before</th>
<th>after</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HU</td>
<td>QU</td>
</tr>
<tr>
<td>Correct</td>
<td>70.9</td>
<td>26.9</td>
</tr>
<tr>
<td>Incorrect</td>
<td>75.8</td>
<td>35.5</td>
</tr>
<tr>
<td>Difference</td>
<td>4.9*</td>
<td>8.6**</td>
</tr>
</tbody>
</table>

**p <0.0001 * p <0.005

not used at both institutions. Before peer instruction, the average response time is 20-30% shorter for correct answers than for incorrect answers (p <0.001). After peer discussion, the average response time is about 10% shorter for correct answers than for incorrect answers (p <0.01).

We also analyzed the average response time difference between incorrect and correct responses \langle t_{incorr} \rangle - \langle t_{corr} \rangle for each individual question. Of the 175 questions posed in this study: 27 had a statistically significant difference (at p <0.05) where \langle t_{incorr} \rangle >\langle t_{corr} \rangle, three questions had a statistically significant difference (at p <0.05) where \langle t_{incorr} \rangle <\langle t_{corr} \rangle and 145 had no significant difference between\langle t_{incorr} \rangle and \langle t_{corr} \rangle. Therefore, for the majority of the questions with a significant time difference, incorrect responses are not misconception-like in nature because students take longer to provide an incorrect answer than the correct one.

Figure 4.1 shows histograms for responses as a function of ConceptTest response time before and after peer discussion. Correct responses (dark grey) and incorrect responses (white) are expressed as a percentage of all responses over all the ConceptTest questions posed during the semester. The light grey area represents the overlap of
Figure 4.1: Distributions of ConcepTest response times before and after peer discussion.

the white and dark grey bars. The top two histograms represent data from Harvard (HU) and the bottom two histograms represent data from Queen’s (QU). Figure 1 illustrates the shorter time scale with which correct responses are given compared to incorrect responses. Given that there was no set amount of time that the polling stayed open, we find some questions were left open for more than 5 minutes.

Figures 4.2-4.5 show the ratios of correct to incorrect responses plotted as a function of the percent of students in the class who have entered their response. Figure 4.6 shows a summary of 4.3-4.5. Ratios of correct to incorrect responses are shown for responses before the discussion (pre) and after the discussion (post) for
Figure 4.2: Ratios of correct to incorrect responses plotted as a function of the percent of students in the class who have entered their response, before discussion in the Harvard classroom. Green bars show the percent of responses that are correct, white bars show the percent of responses that are incorrect. Error bars show the standard error of the mean.
Figure 4.3: Ratios of correct to incorrect responses plotted as a function of the percent of students in the class who have entered their response, after discussion in the Harvard classroom. Green bars show the percent of responses that are correct, white bars show the percent of responses that are incorrect. Error bars show the standard error of the mean.
Figure 4.4: Ratios of correct to incorrect responses plotted as a function of the percent of students in the class who have entered their response, before discussion in the Queen’s classroom. Green bars show the percent of responses that are correct, white bars show the percent of responses that are incorrect. Error bars show the standard error of the mean.
Figure 4.5: Ratios of correct to incorrect responses plotted as a function of the percent of students in the class who have entered their response, after discussion in the Queen’s classroom. Green bars show the percent of responses that are correct, white bars show the percent of responses that are incorrect. Error bars show the standard error of the mean.
Figure 4.6: Ratio of correct to incorrect responses as a function of the percentage of responses received.

Figure 4.6 shows that as more students respond, the proportion of correct answers decreases compared to the proportion of incorrect answers, both before and after the peer discussion. Once about 80% of the students in the class have responded, the ratio of correct to incorrect answers rapidly approaches levels indicating random guessing by the students.

Table 4.2 displays the average response times for correct and incorrect answers for questions determined to be easy and hard. The difference in response time was
Table 4.2: Average time taken for correct and incorrect answers on ConceptTests classified as easy and hard, before and after peer discussion

<table>
<thead>
<tr>
<th></th>
<th>HU (s)</th>
<th>QU (s)</th>
<th>HU (s)</th>
<th>QU (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>before</td>
<td>after</td>
<td>before</td>
<td>after</td>
</tr>
<tr>
<td>Correct</td>
<td>52.9</td>
<td>65.0</td>
<td>24.9</td>
<td>35.3</td>
</tr>
<tr>
<td>Incorrect</td>
<td>62.2</td>
<td>69.9</td>
<td>29.1</td>
<td>36.1</td>
</tr>
<tr>
<td>Difference</td>
<td>9.3**</td>
<td>4.9</td>
<td>4.2**</td>
<td>0.8</td>
</tr>
</tbody>
</table>

* **p <0.0051, * p <0.005

Determined by subtracting the average response time for correct answers from the average response time for incorrect answers. The difference is positive for easy and hard questions both before and after the peer discussion indicating that, regardless of the difficulty of the question or whether the response is provided before or after the discussion, students take longer to respond with an incorrect answer than with a correct one. However, this response time difference is only statistically significant for easy questions (p <0.0001). For difficult questions, while the average response time is longer for incorrect answers than for correct answers, the difference is small and not statistically significant. Table 4.2 also shows that students take longer to answer hard questions than easy questions, regardless of whether they are answering correctly or incorrectly.

Figure 4.7 shows the response time before and after the discussion for responses that fall into each of the five transition categories, discussed in chapter 3. Responses that are right in both rounds are made, on average, significantly faster both before and after the discussion (p<0.005). Responses that are right in round 1 and go
Figure 4.7: Average response time before and after peer discussion for questions that are answered correctly in both rounds (RR), switched from right to wrong (RW), switched from wrong to a different wrong (WW-D), switched from wrong to right (WR) and answered with the same incorrect answer in both rounds (WW-S).

Figure 4.7: Average response time before and after peer discussion for questions that are answered correctly in both rounds (RR), switched from right to wrong (RW), switched from wrong to a different wrong (WW-D), switched from wrong to right (WR) and answered with the same incorrect answer in both rounds (WW-S).

...on to be switched to a wrong answer in the second round are made, on average, significantly slower both before and after the discussion. Aside from right to right, which is significantly faster than all the other transitions, there is no statistically significant difference in response times for any of the other transitions, before or after the discussion.

Figure 4.8 shows that students with more incoming physics knowledge and higher self-efficacy respond faster, both before and after the peer discussion. Figure
Chapter 4: Conceptual question response times in Peer Instruction classrooms

Table 4.3: Regression models predicting student response times for correct and incorrect responses both before and after the discussion

<table>
<thead>
<tr>
<th></th>
<th>Before Discussion</th>
<th></th>
<th></th>
<th>After Discussion</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Incorrect</td>
<td>Correct</td>
<td>Incorrect</td>
<td>Correct</td>
<td>Incorrect</td>
</tr>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
<td>Model 1</td>
<td>Model 2</td>
<td>Model 1</td>
<td>Model 2</td>
</tr>
<tr>
<td>Gender</td>
<td>9.5*</td>
<td>2.38</td>
<td>3.71</td>
<td>1.16</td>
<td>8.5*</td>
<td>4.9</td>
</tr>
<tr>
<td>Pre-course physics knowledge</td>
<td>-1.27***</td>
<td>-0.67*</td>
<td>-1.14***</td>
<td>-1.14***</td>
<td>0.58***</td>
<td>0.39***</td>
</tr>
<tr>
<td>Pre-course self-efficacy</td>
<td>0.58***</td>
<td>0.50***</td>
<td>0.50***</td>
<td>0.50***</td>
<td>0.39***</td>
<td>0.39***</td>
</tr>
<tr>
<td>R²</td>
<td>0.15</td>
<td>0.14</td>
<td>0.15</td>
<td>0.14</td>
<td>0.14</td>
<td>0.11</td>
</tr>
<tr>
<td>RMSE</td>
<td>62.3**</td>
<td>65.4</td>
<td>55.5</td>
<td>56.7</td>
<td>59.7</td>
<td>59.6</td>
</tr>
</tbody>
</table>

*p < 0.05  ** p < 0.01  *** p < 0.001
Chapter 4: Conceptual question response times in Peer Instruction classrooms

4.8 displays average response times before and after the peer discussion for students in four different groups: (1) below average pre-course CSEM scores, (2) above average pre-course CSEM scores, (3) below average pre-course self-efficacy scores, and (4) above average pre-course self-efficacy scores. Students with higher incoming physics knowledge (above average pre-course CSEM scores) respond approximately 15% ($p < 0.001$) faster before the peer discussion, and around 10% ($p < 0.001$) faster after the peer discussion, than students with lower incoming physics knowledge (below average pre-course CSEM scores). Students with above average scores on the pre-course self-efficacy survey respond about 10% ($p < 0.001$) faster both before and after the peer discussion, than students with below average scores.

Figure 4.9 shows male and female students response times for correct and incorrect answers. The first set of bars in each cluster (model 1) displays the response times when controlling for gender only while, the second set of bars (model 2) displays the response times after adding pre-course CSEM and self-efficacy scores to model 1 as additional predictor variables. The left plot of Figure 4.9 indicates that males respond faster than females before discussion, regardless of whether they have the answer correct or incorrect, but that this difference disappears when students’ pre-course knowledge and self-efficacy are controlled for. The right plot of figure 4.9 shows that the same is true when considering male and female response times after the peer instruction. The regression parameters and significance metrics for both models are displayed in table 4.3.

Table 4.4 shows that a gender difference exists in the number of attempts students register before committing to a final answer.
Figure 4.8: Average response times before and after discussion for students with below and above average pre-course CSEM scores and below and above average pre-course self-efficacy scores.

Table 4.4: Average number of attempts for males and females on ConceptTests, before and after the peer discussion

<table>
<thead>
<tr>
<th></th>
<th>$\langle n_{before} \rangle$</th>
<th>$\langle n_{after} \rangle$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>2.4</td>
<td>3.0</td>
</tr>
<tr>
<td>Females</td>
<td>1.7</td>
<td>1.8</td>
</tr>
<tr>
<td>Difference</td>
<td>1.3**</td>
<td>1.2**</td>
</tr>
</tbody>
</table>

***$p < 0.0001$
Compared to female students, male students respond significantly more times before deciding on a final response. The average number of attempts made by males exceeds that made by females both before \( n_{\text{before}} \) and after \( n_{\text{after}} \) the peer discussion. Before the peer discussion, male students change their initial response 40% more times than female students \( (p < 0.001) \) before deciding on a final response. After the peer discussion, male students change their initial response 65% more times than female students \( (p < 0.001) \) before deciding on a final response.

4.4 Discussion

4.4.1 Incorrect answers take more time

Multiple-choice ConcepTests are designed so that the incorrect choices are distractors, that is, incorrect ideas that are commonly held by students. These conceptions are often portrayed as common-sense beliefs that are stable and resistant to
Chapter 4: Conceptual question response times in Peer Instruction classrooms

instruction [11]. Previous research has shown that when students answer conceptual questions with misconception-like responses, they tend to respond more quickly than those answering correctly [65]. In other words, strong distractors should yield quick responses. However, we find that students take longer to respond when they answer incorrectly, suggesting that these choices are not seen as strong distractors that yield automatic responses. Our data therefore suggests that when students select an incorrect answer, it is more likely because they do not know the answer, rather than because they are confident about the wrong answer.

We find that students take longer to respond to more difficult questions. Instructors frequently adjust their pacing based on the fraction of the class that has responded to a question, leaving the poll open longer for more difficult questions. Therefore, the student response rate for a question dictates the length of time the poll for that question is left open, but only up to a point. Instructors often close the poll before 100% of their class has responded though, according to our data, rarely when less than 77% has voted. On average, in the two classrooms used in this study, polls were closed once 91% of students had responded. Figure 4.6 shows that once more than 80% of students have responded, the ratio of correct to incorrect answers is so low that instructors should consider closing the poll. Once 90% of students have responded, the correct to incorrect ratio is, on average 25%, indicating that students are doing no better than random guessing.

We find that the difference in response time between correct and incorrect answers depends on the difficulty of the question. For easy questions, incorrect answers are not automatic; they take significantly longer than correct answers. Students an-
swering easy questions incorrectly are spending time thinking. For harder questions, this does not appear to be the case. When the questions are difficult, the time taken to give an incorrect answer does not differ statistically from the time taken to give a correct answer. Apparently, the correct choice is not as obvious as for easy questions and students take as much time evaluating correct and incorrect options.

Response time for both correct and incorrect answers is significantly longer after peer discussion. Figure 1 shows that after discussing the question with their peers, students take longer to respond to the question than when they respond individually. There are two possible explanations for this finding. It is possible that response times are longer in the second round \( t_{after} \) because students take more time to think about the question. Alternatively, it could simply be due to the fact that some students may continue talking after the polling has reopened and that \( t_{after} \) includes some of the discussion time.

Our findings are based on CT response time data collected at two different institutions in two different classrooms with instructors who had very different pacing. A comparison of the top two histograms and the bottom two histograms in Figure 1 illustrates the different pace at which CTs were posed to students at Harvard (top) and to students at Queen’s (bottom). On average, the polling in the Harvard classroom was kept open much longer than in the Queen’s classroom, both before and after the discussion. Despite this difference in pacing, the same time scale difference emerges between correct and incorrect responses. Regardless of the institution and how long students are given to answer the questions, incorrect answers take more time than correct answers.
4.4.2 Timing and ConcepTest Switching

In chapter 3 we showed that a significant proportion of CT switching (27%) occurs in directions that are negatively associated with student learning (right to wrong and wrong to different wrong). However, our data on timing, presented in this chapter, also suggests that when students switch in negative directions, they are spending just as long thinking about the question in round 1 and discussing the question in round 2 as when students switch in positive directions. Figure 4 shows that response times, both before and after the peer discussion, do not differ based on switching. While students answer significantly faster when they have the correct answer in both rounds ($p < 0.001$), there is no statistically significant difference in response times (for round 1 or 2) between any of the other transitions. That the time on task during both rounds of questioning during right to wrong and wrong to wrong switching is just as long as for wrong to right switching suggests that the Peer Instruction discussion still caused students to engage with the question to a similar degree, and students exhibiting these types of switching are not just guessing hastily or answering just for the sake of registering a response.

4.4.3 Response times vary according to students’ pre-course characteristics

Figure 4.8 shows that the students who respond faster are those who know more physics and have a stronger belief in their ability to be successful in a physics course at the beginning of the semester. These students spend less time considering alternative answers, that is, their answers are more automatic. Interestingly, these
students answer more quickly, regardless of whether they are answering correctly or incorrectly. This finding suggests that response times are dependent on intrinsic student characteristics as well as cognitive processes.

4.4.4 CTs Responses and Gender

In contrast to earlier findings [67], we find that although males respond significantly faster than females, the difference disappears after controlling for self-efficacy and CSEM scores. The gender difference in response times appears to be at least partly attributable to a difference in pre-course knowledge and self-efficacy. When answering correctly, males respond 20% faster than females before \((p<0.05)\) the discussion and 10% faster after the \((p<0.05)\) discussion. When answering incorrectly, the difference is less pronounced. Before the discussion males answer incorrectly 6% faster than females (albeit this difference is not significant) and 12% faster after the discussion \((p<0.05)\). However, when a linear regression model is used to control for pre-course knowledge and self-efficacy, gender ceases to be a significant predictor of response times and the gender difference disappears completely.

Our results also show that, compared to female students, male students respond significantly more times before deciding on a final response. This result is consistent with previous findings that males are more likely than females to answer with more attempts \([68]\) and change their answer within a single round before committing to a final response \([67]\).
4.5 Conclusion

For two different student populations, we find four interesting results from an analysis of response times to conceptual questions posed in class. The first is that incorrect answers take more time than correct answers, especially for easy questions. This suggests that incorrect responses result from students not knowing the answer, rather than from strongly held misconceptions. The second is that when students switch to a wrong answer, they spend just as long thinking and discussing the question as when students switch to a right answer, suggesting that students benefit from Peer Instruction even when switching to a wrong answer. The third is that students with greater incoming physics knowledge and higher self-efficacy respond faster, indicating that response times are partly a function of student characteristics. Fourth, there is no gender difference in response rate when other student characteristics are controlled for. In light of these findings, we recommend that instructors terminate polls once 80% of the answers are in, because at that point an increasing fraction of students respond by random guessing.
Chapter 5

The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

5.1 Introduction

Discussion forums are an integral part of all online — and many offline — courses, but little is known about how to effectively incorporate them into existing course structure and how to engender meaningful learning and assessment. In this chapter, we explore ways to steer discussion forums to produce high-quality learning interactions. In the context of a flipped, introductory physics class, we investigate three questions: What is the relationship between students’ participation in the online forum and their performance in the course? What effect does seeding the forum with comments from previous iterations of the course have on student participation in the forum? What
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

effect does varying the sizes of sections of students who can see each other’s comments have on student participation in the forum? Before exploring these three questions, we provide context with details on Applied Physics 50, the course in which the study is conducted; and NB, the online discussion forum used.

It is generally accepted that students understand material better after discussing it \[69,70\]. Discussion forums have been used successfully as tools to facilitate interactions and exchanges of knowledge between learners and between learners and instructors \[71,74\]. The social constructive theory of learning with technology by Brown & Campione (1996) emphasizes that successful learning requires continuous conversation between learners as well as between instructors and learners \[75\]. The asynchronous nature of online discussion forums allows for discussion between learners and between learners and instructors at any time of day or night, and this is a major advantage over other forms of communication \[75\]. Some studies have shown that participation in online discussion forums is more active when it is linked to assessment \[73\]. Other studies have shown that assigning a grade for participation on the forum is necessary to ensuring that students take part \[76\]. Despite their increasing use in both online and residential courses, there is little existing work on how participation in online discussion facilitates learning.

5.1.1 Applied Physics 50

Our study is carried out in the context of Applied Physics (AP) 50, an introductory physics class offered by the School of Engineering and Applied Sciences at Harvard. It is a year-long course divided into two parts. Mechanics is the focus of
the first part, taught during the fall semester and electricity and magnetism is taught
during the spring semester. The course has been running for two years, having been
first offered fall, 2012. Between the first and second years, student enrollment tripled
(from 28 students in 2012 to 88 students in 2013). Our study is based on three
semesters worth of data from AP50 (spring 2013, fall 2013 and spring 2014).

Applied Physics 50 is an example of a flipped classroom. There are no lectures,
students are expected to read the textbook online and develop a certain level of
comfort with the material before coming to class. Then, while in class, students work
in small teams on projects and activities that are designed to help them better learn
the material. The pedagogy used in AP50 draws on features from both Project Based
Learning [77] and Team Based Learning [78]. Project Based Learning is a teaching
strategy in which students work for an extended period of time on an inquiry-driven
project, often inspired by a real-world problem. By researching and problem-solving,
students gain knowledge and skills in specific content areas. All of the learning goals
for AP50 are addressed through three projects that students work on in class, as part
of a team. Team Based Learning is a teaching strategy that has students organized
into small, permanent groups. Students work within these groups for all aspects
on the course, including assessments, which are taken together as a team. “The
effectiveness of team learning as an instructional strategy is based on the fact that it
nurturesthe development of high levels of group cohesiveness which, in turn, results in
a wide variety of other positive outcomes.” (Michaelsen, 1994, p.6) Peer Assessment
is an important aspect of Team Based Learning; students evaluate their teammates’
contributions and this assessment contributes significantly to the final grade. At least
part of the success of Team Based Learning is attributable to the fact that students
work harder because they are made to feel accountable to their teammates. Students
in AP50 are organized into teams throughout the semester. For the three projects,
each of which lasts approximately one month, students are working closely with three
or four of their classmates, as part of a team. The teams change for each of the
projects, to allow students the opportunity to work with a variety of their peers and
further develop their interpersonal skills. The class meets twice weekly for 3 hours.
There are no additional sections or labs; all of the course components are contained
within these two, 3 hour blocks. There are six different types of in-class activities,
each of which is described below. In addition to having time to work on the projects,
students use time in class engaging in 1-3 of the activities, which are scheduled into
each of the blocks of class time.

Each of these class activities are designed to help students master the relevant
physics and get started on the projects, which serve as the focal point for the course.
The class activities are interrelated and are scheduled in an order so that they build
on one another and provide scaffolding as students learn the concepts. Each of the
six activities is described below in the order that they appear for each unit.

Learning Catalytics: at the beginning of each unit, the instructor conducts a 90-
minute Peer Instruction session during which students answer ConcepTests on difficul-
t concepts selected from the unit. Students answer individually initially and, after
discussing the question with their team, they answer again. These Learning Catalyt-
ics sessions are placed at the beginning of the unit as they allow the instructor to
probe students’ understanding of the reading and address difficult concepts. Univer-
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

University of Washington Tutorials: are worksheets that are designed to address common misconceptions about the course content. During this activity, which lasts an hour, students work with their team on the worksheet and this allows them to explore their thinking about the more difficult concepts in the material.

*Estimation Activity:* is a list of five unknown quantities that are related to the content of the unit. Students are given thirty minutes to think and work with their team to estimate the quantities to the nearest order of magnitude.

*Experimental Design Activity:* are online (typically PhET) simulations or hands-on, lab-like activities that help students develop experimental and analytical skills that are important for the projects.

*Problem Set Reflection:* the problem sets are comprised of 4-5 traditional physics problems that students are given a week to work on at home. Students are instructed to give the problems their best effort before coming to class and to bring their solutions to class to work through with their team, during the problem set reflection activity. During this time students work with their teams to discuss and improve their solutions, resolve conceptual difficulties and reflect on areas that need to be reviewed. At the end of this activity, students submit their revised solutions with a written reflection on the aspects of the problem set that they struggled with and how they resolved misunderstandings.

*Readiness Assurance Activity (RAA):* is the closest thing to a traditional exam in AP50. RAAs are assessments conducted in-class, at the end of each unit, to assure that everyone is on track in the learning of the basic concepts. During the first half of each RAA, students work individually to answers a set of problems. Students are free
to consult the textbook or the internet but are not allowed to discuss the problems with one another. During the second half of each RAA, students get together with their team and discuss the same problems, towards the goal of agreeing on the correct answer. After reaching a consensus, each team must submit a final answer to each question, as a group. After submitting their team round response, the correct answer is revealed to the group. This second part of the RAA provides an opportunity to learn collaboratively in teams as well as receive immediate feedback. The overall RAA score is determined by a combination of the student’s individual score (50%) and their team’s score (50%).

5.1.2 NB

As AP50 has no lectures, the reading is the primary exposure to the content of the class. Before each class, students are required to read and annotate a specific chapter in the textbook. Students access the textbook via NB, an online collaborative textbook annotation tool developed by the Computer Science & Artificial Intelligence Lab at the Massachusetts Institute of Technology. NB is a forum that is somewhat unusual for being situated in the margins of the (online) course textbook. Like most forums, NB supports threaded discussions of the material; placement in the margins simply improves the organization of the discussions in context of the textbook material.

The textbook is uploaded to the NB website, to which students log on to read and annotate the text. Annotations are made by highlighting a passage of the textbook and typing into a text field that appears in the margin. Students annotate
by asking questions about what they are reading and also by responding to other questions and comments made by their classmates. Annotations are organized into ‘threads’, which constitute a starting comment or question followed by all the replies made by other students to the initial annotation or to the subsequent replies. In this way, students have a discussion about specific aspects of the content, within the context of the textbook. Due to the flipped nature of AP50, NB is the only mechanism for delivering the course content to the students, entirely supplanting traditional lectures. It is therefore very important that students read thoughtfully before coming to class, to be prepared to participate and learn from the in-class activities and from each other.

5.1.3 NB Assessment

Given the important role that NB plays in the structure of AP50, students are assessed on the thoughtfulness of their annotations and this assessment is important in determining each student’s final grade. The course is divided into five units, each of which constitutes 3-4 chapters of the textbook. Each class session pertains to a single chapter in the textbook which students are expected to read beforehand. To encourage students to stay on top of the reading, their annotations are evaluated at the end of each unit and this represents a component of the overall course grade. Annotations are evaluated on the basis of both quality and timeliness. The criteria used to assess quality and timeliness was modified between the first and second year of the course.
Timeliness

In the first year, annotation deadlines were not strictly enforced and we gave students credit for any annotation entered into NB within 7 days of the class that pertained to that chapter. Students had a week after the class that dealt with the chapter to annotate it. While the instructor emphasized the importance of being prepared for the class activities by reading and annotating the chapter beforehand, students were not penalized for coming to class without having read the chapter. As the students were engaging in all classroom activities as part of their teams, we prescribed to the Team Based Learning philosophy that students would feel accountable to their teammates and this would be incentive enough to read the textbook before coming to class.

In the second year, we adopted stricter criteria with which to evaluate the timeliness of the NB annotations. Full credit was only given to annotations entered before the class pertaining to that chapter. Annotations submitted after the relevant class but before the end of the unit, only received half credit. To encourage students to continue discussing the material after class, we extended each of these two deadlines by three days for students who were replying to other students’ annotations. Therefore, replies to annotations were assigned full credit as long as they were entered within three days of the relevant class.

Quality

In the first year, we assessed the quality of the annotations with a rubric adapted from the Trends in International Mathematics and Science Study (TIMSS) Assess-
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

Framework [80]. Each annotation was assessed on a four-point scale based on which cognitive domain it most closely exemplified (Knowing, Applying, Reasoning). A summary of the rubric used to evaluate the quality of annotations in Year 1 of the course is outlined in table 5.1.

While the instructors informed students that the quality of their annotations was being assessed, the Year 1 rubric was not distributed to the students and so, there was very little transparency in the assessment of the annotations.

In the second year, to improve the issue of low-grade transparency from the first year, we distributed a rubric outlining how annotations would be evaluated, both for timeliness and quality. We also provided a series of sample annotations from the first iteration of the course, with a key and explanation of how each would be evaluated. The sample annotations can be found in Appendix 3 and the Year 2 quality rubric is outlined in table 5.2.

In an attempt to simplify the Year 1 rubric, we made several significant modifications. We changed the scale from four points to three points and we shifted the focus from cognitive domains, which we thought were too steeped in educational jargon for students to understand, to an emphasis on demonstrating thorough and thoughtful reading of the text. To encourage quality over quantity, there was no requirement for a specific number of annotations per chapter. To provide students with a rough guideline for quantity, we included the following statement in the Year 2 syllabus: “It is unlikely that effort will be reflected by just one or two annotations per chapter, unless your annotations are unusually thoughtful and stimulate a deep discussion. On the other extreme, 30 per chapter is probably too many, unless they are
Table 5.1: Year 1 rubric for evaluating the quality of NB annotations

<table>
<thead>
<tr>
<th>Score</th>
<th>Description &amp; Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (no physics)</td>
<td>Annotation devoid of physics content</td>
</tr>
<tr>
<td>1 (knowing)</td>
<td>Demonstrates surface level reading of the textbook. These are reading comprehension annotations, which do not exhibit any interpretation, application, analysis, reasoning or synthesis of the content. A question about the definition of a term or an explanation that makes a claim without any substantiation are examples of the types of annotations that fall into this category.</td>
</tr>
<tr>
<td>2 (applying)</td>
<td>Demonstrates some interpretation of the textbook. These are annotations that compare, contrast, and classify; interpret scientific information in light of a science concept or principle; apply an understanding of science concepts and principles to find a solution or develop an explanation. Questions identify specifically where the misunderstanding is and explanations substantiate claims with facts.</td>
</tr>
<tr>
<td>3 (reasoning)</td>
<td>Demonstrates consideration of a number of different factors or related concepts; make associations or connections between concepts from different chapters; demonstrate understanding of unified concepts and themes; integrate mathematical concepts or procedures. Questions focus on the deeper connections between concepts. Explanations substantiate claims with reasoning and theoretical assumptions</td>
</tr>
</tbody>
</table>
very superficial comments or questions. Somewhere in between these two extremes is about right. Last year the mean number of annotations per student was about 10/chapter.”

During the first year of AP50, a class of thirty students generated more than 3500 annotations in a single semester. In the second year, when the class grew to 90 students, grading the annotations by hand became impossible. To deal with this issue of scaling, we developed an automated annotation scoring algorithm using a data set of 2000 annotations from the previous year. We extracted features such as total word count, parts of speech, unigrams, stem n-grams and punctuation from the training set of annotations. We used a linear regression model to learn from these features and generate parameters for testing and validation. We used a forward feature selection algorithm to arrive at a combination of features that provides the best score prediction. We used Cohen’s Kappa to measure the inter-rater reliability
between human graders and between human graders and the grade predicted by the model. Our final model was able to achieve a Kappa of 0.74, which surpassed the average kappa between two human graders (0.60). We have been using this automated algorithm to machine grade the quality of student annotations in AP50 since October 2013.

5.1.4 Final Grade

The final grade for AP50 is determined based on students’ performance on four course components: contributions to the NB system assessed on the quality and timeliness of annotations, problem sets assessed on individual effort and the quality of a self-evaluation, Readiness Assurance Assessments (RAAs) determined by both individual and team scores, and projects assessed on meeting the project criteria. During the first three semesters of AP50 (fall 2012, spring 2013, fall 2013), a traditional algorithm was used to arrive at a final grade based on the four course components. During the first two semesters, NB contributions represented 10%, problem sets 20%, RAAs 30%, and projects 40% of the final grade. In the fall of 2013 this algorithm was modified slightly to increase the weighting of the NB annotations from 10% to 15% and the weighting of the problem sets from 20% to 25%. RAA’s remained at 30% while projects were reduced to 30%. These modifications were made to emphasize the importance of the components of the course that involved self-directed learning (namely the NB annotations and the problem sets).

In the fourth semester of AP50 (spring 2014), the traditional final grade algorithm was replaced with a domain-based approach. Students were assessed, on a
Table 5.3: Course component factors broken down into the four domains used to calculate final course grade

<table>
<thead>
<tr>
<th>Domain</th>
<th>Contributing Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-directed learning</td>
<td>NB annotations and Problem Sets</td>
</tr>
<tr>
<td>Achievement of learning goals</td>
<td>RAAs and Project Report</td>
</tr>
<tr>
<td>Team Work</td>
<td>Project Presentation and Peer Assessment</td>
</tr>
<tr>
<td>Professionalism</td>
<td>Participation, Punctuality, and Engagement</td>
</tr>
</tbody>
</table>

three-point scale, on four separate domains. The same course components described above were organized into each of these four domains (in addition to some other course components, not mentioned above). A description how the course components factor into each of the four domains can be found in table 5.3.

The three-point scale used to assess each of the following domains is as follows:

3 = significantly exceeds expectations (given only in the most exceptional cases)
2 = meets expectations
1 = improvement needed
0 = deficient

The computation of the final grade is based on a combination of the scores in each of the four domains such that:

A = 2 (or higher) in each of the four domains
A- = three 2, one 1
B+ = two 2, two 1
5.2 Methods

We are interested in understanding the role that online discussion forums play in student learning, specifically in the context of flipped classrooms. Towards this goal, we investigate three research questions: What is the relationship between students’ NB participation and their learning in the course? What effect does seeding the forum with comments from previous iterations of the course have on student participation in the forum? What effect does varying the sizes of sections of students who can see each other’s comments have on student participation in the forum?

5.2.1 Sectioning

To explore the effect of sectioning, in the fall of 2013, we started subdividing the class into smaller and smaller online sections as the semester progressed. We hypothesize that there is a “sweet spot” for sizing discussion forums. We speculate that when the discussion group is too large, students will find they have nothing left to say, while if it is too small, interesting discussion topics may be missed. At the end of the first unit, students were randomly sorted into two sections. Instead of
all students being able to see and respond to all annotations class-wide, they could only see and respond to the annotations from the students in their section. This subdivision effectively decreased the online forum size from 88 students to 44 students. At the end of the fourth unit, we divided the class further into four randomly sorted groups, each containing 22 students. For the final unit, we divided the class even further into eight sections of 11 students.

The following semester, in spring 2014, we divided the class into four sections with approximately 20 students in each and kept these sections constant throughout the entire semester. Throughout the two semesters, we randomly assigned students to the sections but made sure that students in each section had the same average pre-course CSEM score. We compare the average thread length as a function of section size to determine if, beyond a certain section size, forums become saturated.

5.2.2 Seeding

We hypothesize that a way to stimulate discussion is to seed the discussion forum with provocative topics, and that a good way to choose these topics is to select them from among the more successful forum discussions from last year’s class. At the beginning of the third unit in fall 2013, we started seeding the discussion forum in this way. We used three different selection criteria to choose the seeded annotations. These criteria are summarized in table 5.4. For seeding condition 1, we selected the top 10 longest threads in their entirety. For seeding condition 2, we selected only the first annotation from the top 10 longest threads, speculating that they might trigger long discussion again. For seeding condition 3, we chose the 10 highest quality
Table 5.4: A summary of the different seeding conditions used in the experiment

<table>
<thead>
<tr>
<th>Seeding Condition</th>
<th>Selection Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Longest threads</td>
<td>top 10 longest threads</td>
</tr>
<tr>
<td>2. First annotations (longest threads)</td>
<td>first annotation from top 10 longest threads</td>
</tr>
<tr>
<td>3. Quality annotations</td>
<td>top 10 highest quality first annotation</td>
</tr>
<tr>
<td>4. No seed</td>
<td>none</td>
</tr>
</tbody>
</table>

annotations that started a thread last year, based on the quality rubric described above. Each seeded thread or annotation was imported from last year’s forum and entered into this year’s forum anonymously. It is common practice for real students in the class to post annotations that are anonymous to everyone except the instructor and therefore, these seeded annotations appear no different from ‘real’ annotations.

In fall 2013, we applied one of the four seeding conditions to each of the NB sections for each unit. We rotated the four seeding conditions through the various sections for each unit. In spring 2014, we adopted a more systematic approach to seeding and sectioning. From the beginning of the semester, students were a member of one of four stable sections of 16-20 students. Unlike fall 2013, in spring 2014, we did not vary seeding conditions over the course of the semester. We seeded two of the four sections with the top 20 longest threads from the previous year (condition 1, except twice the number of threads). The other two sections we did not seed at all (condition 4).
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

5.2.3 Student Learning Metrics

To study the relationship between students’ NB participation and their learning, we use a number of course metrics. In addition to final grade, we measure learning with conceptual surveys, ConcepTests and RAA scores. At the beginning and the end of each semester, we administered a conceptual survey as a pre and post-test. As the spring semester covers electricity and magnetism, we used the Conceptual Survey on Electricity and Magnetism (CSEM) [10] in the spring of 2013 and the spring of 2014. The fall course covers mechanics and so, in the fall of 2013, we used the Force Concept Inventory (FCI) [11]. Both surveys are validated instruments widely used to measure students’ conceptual understanding of Newtonian dynamics (in the case of the FCI) and basic principles of electricity and magnetism (in the case of the CSEM). In addition to collecting students’ scores to these pre- and post-tests, we also collect their individual responses to ConcepTests posed during the Learning Catalytics activities, over the course of the semester. We also use students’ individual (round 1) responses to the Readiness Assurance Assessments as another measure of learning in the course.

5.2.4 Qualitative Annotation Metrics

In addition to assessing the quality of the annotations, we classify the NB annotations on a number of other qualitative dimensions to be able to measure students’ level of engagement on the online forum. The two rubrics used to assess the quality have already been discussed. We used the quality classification both for assessment of in the context of the course as well as for research purposes. We developed addi-
tional coding schemes for classifying annotations and threads on a number of other dimensions, for research purposes only. Most broadly, all annotations are classified as one of three types: comments, questions, and explanations. Comments are always the first annotation in a thread and are statements about the textbook that are made without the expectation of a reply. Questions, on the other hand, are posed with the expectation of a response with information or an explanation of some concept. Questions can either be the first annotation in a thread, or not. Explanations are always written in response to a question and are therefore never the first annotation in a thread. Explanations are classified further with a framework developed to measure the level of argumentation provided through the explanation [81]. This framework ranks explanations based on the presence of claims, data, warrants, backing and qualifiers. A claim is a conclusion whose merits need to be supported with data, warrants and backing. Data are facts used to support the claim and warrants are reasons that establish the connections between the data and the claim. Backing is a theoretical assumption on which a warrant rests. High-level explanations support claims with data, warrants and backings. The highest level explanations include qualifiers, which establish the boundaries of the claim. Low level explanations make claims with no warrants or backing. We used this framework to rank explanations on a five-point scale described in table 5.5.

We coded discussion threads using an adapted scheme developed to examine discourse patterns and collaborative scientific reasoning in peer discussions [82]. We define a thread as being made up of at least two annotations (i.e. we ignore isolated annotations). We categorized threads as one of the following types: consensual, re-
Table 5.5: Framework used to code students’ explanations

<table>
<thead>
<tr>
<th>Score</th>
<th>Explanation: Description and Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Explanation makes a claim and clearly justifies it with a warrant and/or backing. The limitations of the claim are described with a qualifier.</td>
</tr>
<tr>
<td>4</td>
<td>Explanation makes a claim and clearly justifies it with a both warrants AND backing.</td>
</tr>
<tr>
<td>3</td>
<td>Explanation makes a claim and clearly justifies it either warrants OR backing.</td>
</tr>
<tr>
<td>2</td>
<td>Explanation makes a claim but does not justify it with warrants or backing.</td>
</tr>
<tr>
<td>1</td>
<td>Explanation makes no claim.</td>
</tr>
</tbody>
</table>

sponsive, transfer, generative and argumentative. The generative and argumentative threads are of particular interest to us as these are the types of activities that research has shown to be most effective in promoting learning. A brief description of each of these categories is summarized in table 5.6 [82].

5.2.5 Qualitative Coding

The coding of the annotations has taken place in stages. During the first semester of the study (spring, 2013) we coded for quality, annotation type, level of explanation and thread type using three human coders. The coders initially coded sets of 100 annotations and compared codes for the purpose of calibration. Once an inter-rater reliability of 70% was reached, the remaining annotations were divided up
Table 5.6: Framework used to code NB threads

<table>
<thead>
<tr>
<th>Thread Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensual</td>
<td>Only one student contributes substantive statements. Other student responds by passive agreement.</td>
</tr>
<tr>
<td>Responsive</td>
<td>Both questions and responses of at least two students contribute to a substantive discussion.</td>
</tr>
<tr>
<td>Transfer</td>
<td>Knowledge is shared in what is typically a longer discussion. No new ideas emerge.</td>
</tr>
<tr>
<td>Generative</td>
<td>New ideas are linked to someone else’s idea and knowledge emerges in a constructivist manner.</td>
</tr>
<tr>
<td>Argumentative</td>
<td>Critical discussion during which there is disagreement between participants.</td>
</tr>
</tbody>
</table>
and coded by the three coders. During the subsequent two semesters of the study, (fall 2013, and spring 2014) we coded annotations for type and quality using the automatic grading algorithm described earlier. We only coded annotations for the level of explanation and thread during the first semester of the study.

5.2.6 Quantitative NB Metrics

In addition to the qualitative coding of the annotations, we also measure students’ level of engagement on NB with several quantitative metrics related to reading and annotating behavior. We have mined the NB system for information about students’ reading behavior (how long they spend reading per chapter, how far into the chapter they read and when, relative to class they read) and annotating behavior (number of annotations, length of annotations and when, relative to class they annotate).

To investigate the relationship between students’ participation on NB and their learning in the course we compare the qualitative and quantitative NB metrics described above to course metrics such as normalized gain, final grade and performance on in-class activities such as Learning Catalytics and Readiness Assurance Assessments. Before exploring this question, we provide an exploratory overview of how and how much students use NB in the context of this flipped introductory classroom. More specifically, we look at:

1) When, relative to the deadline, are students reading and annotating the textbook.
2) How much time students spend reading the textbook.
3) How far into each chapter students typically read.
4) How much students annotate (per chapter).
5) The proportion of annotations that are comments, questions and explanations.
6) The distribution of annotation quality.
7) The distribution of thread type.

5.3 NB Use

5.3.1 Results
When, relative to the deadline, are students reading and annotating the textbook?

Figures 5.1 - 5.2 are histograms of active reading sessions relative to the class that is relevant for each chapter. Each sub-plot represents a single book chapter and the horizontal axis is time (in hours) relative to the class when the reading is covered. Zero on the horizontal axis represents the time when the class occurs and the reading is due. An active reading session is defined as a period of time students spend actively scrolling on a single page of the textbook that is greater than 30 seconds and less than an hour. Sessions less than 30 seconds are not counted because they are assumed to be students just checking something as opposed to actually reading. Sessions longer than an hour are also eliminated because it is unlikely that a student would spend greater than an hour on a single page (this is probably a case of having another browser window open while being logged into NB). Figure 5.1 shows the timing of the active reading sessions for spring 2013 and figure 5.2 shows the timing of the
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

Figure 5.1: Histograms of when students are reading relative to class time for spring 2013. Time has been normalized so that zero represents the class time for the class when that chapter is being covered.

active reading sessions for fall 2013 (chapters 1-17) and spring 2014 (chapters 22-34). Figures 5.3 - 5.4 are histograms that show when students are annotating relative to each chapter. Again, zero on the x-axis is the time when the class occurs. The black vertical reference lines in figure 5.3 represents the annotation deadline for each chapter in spring 2013. As was mentioned in the timeliness section, the annotation deadline in spring 2013 was around a week after the relevant class (although, for many chapters, that deadline was extended 2-3 days). Figure 5.3 does not show the same vertical reference lines because the annotation deadline in these two semesters was the same as the class time (zero on the x-axis).
Figure 5.2: Histograms of when students are reading relative to class time for fall 2013 & spring 2014. Time has been normalized so that zero represents the class time for the class when that chapter is being covered.
Figure 5.3: Histograms of when students are annotating relative to class time for spring 2013. Time has been normalized so that zero represents the class time for the class when that chapter is being covered.
Figure 5.4: Histograms of when students are annotating relative to class time for fall 2013 & spring 2013. Time has been normalized so that zero represents the class time for the class when that chapter is being covered.
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

Table 5.7: mean (SD) number of hours before (-) or after (+) class students read and annotate for each of the three semesters (spring 2013, fall 2013 and spring 2014)

<table>
<thead>
<tr>
<th></th>
<th>mean (SD) hours relative to class</th>
<th>mean (SD) hours relative to class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>reading</td>
<td>annotating</td>
</tr>
<tr>
<td>spring 2013</td>
<td>163.5 (166.0)</td>
<td>148.8 (70.6)</td>
</tr>
<tr>
<td>fall 2013</td>
<td>47.3 (236.6)</td>
<td>-17.5 (77.5)</td>
</tr>
<tr>
<td>spring 2014</td>
<td>70.2 (259.2)</td>
<td>-17.7 (57.9)</td>
</tr>
</tbody>
</table>

Table 5.7 shows the mean (and standard deviation) number of hours before or after class that students read and annotate for each of the three semesters (spring 2013, fall 2013 and spring 2014). Negative numbers mean that, on average, students are reading/annotating before class whereas positive numbers mean that, on average students are reading/annotating after class.

How much time students spend reading the textbook

Table 5.8 shows the mean (and standard deviation) amount of time students read for each of the three semesters (spring 2013, fall 2013 and spring 2014). The first column of table 5.8 shows the mean number of hours students read per chapter, the second column shows the mean number of active reading sessions per chapter and the third column shows the mean number of minutes per active reading session.
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

Table 5.8: mean (SD) amount of time students read for each of the three semesters (spring 2013, fall 2013 and spring 2014)

<table>
<thead>
<tr>
<th></th>
<th>mean (SD)</th>
<th>mean (SD)</th>
<th>mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>number of hours</td>
<td>number of reading sessions</td>
<td>number of minutes</td>
</tr>
<tr>
<td></td>
<td>per chapter</td>
<td>per chapter</td>
<td>per session</td>
</tr>
<tr>
<td>spring 2013</td>
<td>2.3 (1.3)</td>
<td>43.2 (26.0)</td>
<td>3.3 (6.71)</td>
</tr>
<tr>
<td>fall 2013</td>
<td>2.3 (2.1)</td>
<td>45.8 (32.7)</td>
<td>3.0 (6.51)</td>
</tr>
<tr>
<td>spring 2014</td>
<td>2.2 (2.4)</td>
<td>42.8 (32.4)</td>
<td>3.4 (7.1)</td>
</tr>
</tbody>
</table>

How far into each chapter students typically read

Figures 5.5-5.7 are histograms that show the number of pages into each chapter students are actively reading. The first page in each chapter was renormalized to be counted as the 0th page and the x-axis represents the number of pages, into each chapter, each reading session occurs.

Table 5.9 shows the mean and maximum number of pages into the chapters that students read and annotate for the 3 semesters (spring 2013, fall 2013, and spring 2014).

How much students annotate

Table 5.10 shows the mean number of annotations per student: per chapter and per semester (spring 2013, fall 2013 and spring 2014). The first column shows the mean number of annotations per student, per chapter and the second column shows
Figure 5.5: Histograms of how far into the chapter students are reading relative to class time for spring 2013. The first page in each chapter is counted as 0 and the x-axis represents the number of pages, into a chapter, each reading session occurs.

Table 5.9: mean (SD) and max (SD) number of pages into the chapter reading occurs (spring 2013, fall 2013 and spring 2014)

<table>
<thead>
<tr>
<th></th>
<th>mean (SD) number of pages</th>
<th>mean (SD) number of pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>reading</td>
<td></td>
<td>annotating</td>
</tr>
<tr>
<td>spring 2013</td>
<td>10.0 (4.9)</td>
<td>25.4 (10.5)</td>
</tr>
<tr>
<td>fall 2013</td>
<td>10.6 (3.5)</td>
<td>24.8 (7.3)</td>
</tr>
<tr>
<td>spring 2014</td>
<td>10.5 (4.2)</td>
<td>25.3 (6.6)</td>
</tr>
</tbody>
</table>
Figure 5.6: Histograms of how far into the chapter students are reading relative to class time for fall 2013. The first page in each chapter is counted as 0 and the x-axis represents the number of pages, into a chapter, each reading session occurs.
Figure 5.7: Histograms of how far into the chapter students are reading relative to class time for spring 2014. The first page in each chapter is counted as 0 and the x-axis represents the number of pages, into a chapter, each reading session occurs.
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

the mean number of annotations per student, per semester.

Table 5.10: mean (SD) number of annotations per chapter and per semester(spring 2013, fall 2013 and spring 2014)

<table>
<thead>
<tr>
<th></th>
<th>mean (SD) number of annotations</th>
<th>mean number of annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>per chapter</td>
<td>per semester</td>
</tr>
<tr>
<td>spring 2013</td>
<td>6.6 (2.8)</td>
<td>86.0 (35.8)</td>
</tr>
<tr>
<td>fall 2013</td>
<td>6.5 (5.9)</td>
<td>113.9 (109.2)</td>
</tr>
<tr>
<td>spring 2014</td>
<td>10.2 (15.1)</td>
<td>153.5 (137.3)</td>
</tr>
</tbody>
</table>

The proportion of annotations that are comments, questions and explanations

Figure 5.8 shows a breakdown of the proportion of annotations that are comments, questions and explanations in spring 2013 (left) and fall 2013/spring 2014 (right). Of all the annotations in spring 2013, 22% were comments, 31% were questions and 47% were explanation. In fall 2013/spring 2014, the ratios are quite different. Of all the annotation in fall 2013/spring 2014, 45% of the annotations were comments, 30% were questions and 25% of the annotations were explanations.

The distribution of annotation quality

Figure 5.9 shows a breakdown of the quality of annotations in spring 2013 (left) and fall 2013/spring 2014 (right). In spring 2013, quality was assessed on a four-point
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

Figure 5.8: A breakdown of the types of annotations in spring 2013 (left) and fall 2013 & spring 2014 (right).

scale (see table 5.1) and in fall 2013/spring 2013, quality was assessed on a three-point scale (see table 5.2).

Figure 5.9: A breakdown of the quality of annotations in spring 2013 (left) and fall 2013 & spring 2014 (right).
Figure 5.10: A breakdown of the type of threads in spring 2013 (left).

The distribution of thread type

Figure 5.10 shows a breakdown of the type of threads in spring 2013, when the threads were manually coded. A description of the different types of threads can be found in table 5.6. The most frequent thread type was the responsive thread (46% of threads). In responsive threads, at least two students engage in a substantive discussion whereby one student in responding to the comments or questions of another student.
5.3.2 Discussion

When, relative to the class, are students reading and annotating the textbook?

A comparison of the peaks of the histograms in spring 2013 (figure 5.1) to the peaks of the histograms in fall 2013/spring 2014 (figure 5.2) illustrates the effects of changing the deadline policy between spring 2013 and fall 2013 on when students read, relative to the class time. In spring 2013, when the annotation deadline was laxer, students read, on average, much later than in the subsequent two semesters, when the annotation deadline policy was stricter. A similar comparison of the peaks of the histograms in spring 2013 (figure 5.3) to the histograms in fall 2013/spring 2014 (figure 5.4) shows that students annotate, on average, much later when the annotation deadline in lax compared to when the deadline is stricter. Figures 5.3 and 5.4 show that the vast majority of annotations come in just before the annotation deadline, regardless of when it is. The deadline in spring 2013 was around one week after the relevant class and was extended a few days beyond that, in many cases. The black vertical lines in figure 5.3 indicate the deadline for each individual chapter. The most popular annotation time (peaks on the histograms) is, in most cases, right before the annotation deadline. In fall 2013/spring 2014 the annotation deadline was the class time (t=0) and again, the peaks on the histograms correspond to the this time, indicating most students annotate right before the deadline.

Table 5.7 shows that in spring 2013, students read, on average, 116 hours later than when students read, on average, in fall 2013 and 93 hours later than when students read, on average, in spring 2014. This table also shows that in spring 2013,
students annotate, on average, 166 hours later than when students annotate, on average in fall 2013 and spring 2014. As the students are evaluated on their annotations, the annotation deadline drive students’ reading behavior. When the annotation deadline is later, as it was in spring 2013, students annotate, and therefore read, later. When the annotation deadline is earlier, as it was in fall 2013/spring 2014, students annotate and therefore, read earlier. Clearly, the change in timeliness criteria between the spring of 2013 and the fall of 2014 effected a change in students’ reading and annotating behavior. When we enforced a firm deadline on the day that the reading was being covered in class, this caused students to read and annotate substantially earlier than when this firm deadline was not in place.

How much time students spend reading the textbook

While changing the annotation deadline affected when students annotated and read relative to the class, it does not appear to affect the amount of time students spend reading the textbook. Table 5.8 shows that students annotate slightly longer than two hours per chapter, regardless of which semester we are looking at. The number of reading sessions per chapter is also constant over the three semesters. On average, students read over 44 separate reading sessions per chapter with each session lasting slightly more than three minutes, which is surprisingly short. This short time suggests that students are often skimming the textbook rather than studying it deeply.
How far into each chapter students typically read

Figures 5.5 - 5.7 show how many pages into each chapter the students read. These figures show that the majority of reading sessions occur in the first few pages of each chapter and that, the frequency of reading sessions drops off quickly as the page number increases. Table 5.9 shows that, on average, the reading sessions occur ten pages into the chapter and that the maximum number of pages student annotate into the chapter is 25 pages. Since each chapter is somewhere between 21 and 44 pages long, students are, on average, only reading between a quarter and a half way into the chapter.

How much students annotate

The number of annotations per student, per chapter stayed constant during the first two semesters, averaging approximately six annotations, per student, per chapter. It is surprising that the amount of annotations did not increase between the first and second semesters given that, the expectations were much better defined in fall 2013 than they had been in spring 2013 and annotations went from being worth 10% of the final grade in spring 2013 to 15% in fall 2013. The number of annotations increased in spring 2014, albeit only marginally, despite the fact that there was no change in the expectations or grading policies. The one significant change in spring 2014, compared to the first two semesters, was students were grouped into stable, small sections for the duration of the semester. Small sections helps prevent forum saturation and this may have influenced the number of annotations made per student. This will be discussed further, below.
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

The proportion of annotations that are comments, questions and explanations

In spring 2013, almost half of the annotations were classified as explanations — students responding to other students’ questions. Slightly more than 20% of annotations were classified as comments, with the remaining 30% classified as questions. In fall 2013/spring 2014, the proportion of questions remained constant but, the proportion of explanations dropped to 25% while the proportion of comments increased to 45%. This reversal of the ratio of comments to explanations is curious given the change in the deadline structure between spring 2013 and fall 2013. Starting in fall 2013, students were given an extra three days to provide explanations to one another, and still receive full credit. In spring 2013, no distinction was made in the timeliness criteria between comments, questions and explanations. One possible explanation for the decrease in explanations in the second and third semesters is that many students did not fully understand the timeliness rubric and did not realize that they would still receive full credit for an explanation provided three days past the regular deadline. Worried about being submitting their annotations on time, students might have posted more comments instead of explanations because, at the time that they annotated, there were not a lot of questions to respond to. Students did annotate, on average, significantly earlier in fall 2013/spring 2014 than they did in spring 2013. Students who post earlier are more likely to post comments and questions than explanations because, the earlier students annotate, the fewer questions there are to respond to.
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

The distribution of annotation quality

Given the change from a four-point scale to a three-point scale, we cannot make a direct comparison of the annotation quality between the spring of 2013 and the fall of 2013. The average annotation quality in fall of 2013 (1.62) was very close to that in spring 2014 (1.65) and therefore the data for those two semesters is presented on the same plot. Relative to the maximum of the scale, the average annotation quality in spring 2013 (1.91) is lower than the average annotation quality in fall 2013/spring 2014. This might be due to the coarseness of the three-point scale (compared to the four-point scale) or because, the expectations were more transparent in fall 2013/spring 2014 than they were in spring 2013.

The distribution of thread type

The majority of threads fall into the responsive category where, most typically, one student asks a question and another student responds to it with factual information. Ideally, the threads are elaborative in nature, with students sharing knowledge or constructing new knowledge through longer discussions. The elaborative, argumentative threads are considered the most valuable for learning as they demonstrate critical thinking through discussing and debating the material. These latter thread types were the most rare in the forum. The majority of threads were factual exchanges of information (responsive threads). One of the questions we explore in a subsequent section is how seeding the forum changes the distribution of thread types.
5.4 The relationship between students’ participation on NB and their learning in the course

5.4.1 Results

Table 5.11 shows the correlations between student learning metrics (RAA performance, Learning Catalytics performance, CSEM gain, and final grade) and both quantitative NB metrics (amount of time spent reading, number of reading sessions, number of annotations made, timing of the reading relative to class) and qualitative NB metrics (average annotation quality and the proportion of annotations that are explanations). The only learning metric that is significantly correlated with any of the quantitative NB is final grade. Students who spend more time reading and students who annotate more have higher final grades than students who spend less time reading and students who annotate less.

The qualitative NB metrics correlate more strongly with the learning metrics (RAA performance, Learning Catalytics and CSEM gain). Students who write high quality annotations do better on Readiness Assurance Activities and have higher normalized gain scores on the CSEM than students who write low quality annotations. The correlation between annotation quality and RAA performance is even significant when controlling for pre-course physics knowledge. The average quality of a students’ annotations is still a significant predictor of how they do on the RAAs (p<0.05), even after using students’ pre-course CSEM scores as a covariate. Students with a high ratio of explanations to annotations also do better on Readiness Assurance Activities and on the Learning Catalytics questions than students with a low ratio
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

of explanations to annotations. Again, this is still the case even after controlling for students’ pre-course CSEM scores.

Table 5.11: correlations between learning metrics and NB metrics

<table>
<thead>
<tr>
<th></th>
<th>RAA</th>
<th>LC</th>
<th>CSEM gain</th>
<th>final grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>amount of time spent reading</td>
<td>0.08</td>
<td>0.07</td>
<td>0.20</td>
<td>0.42*</td>
</tr>
<tr>
<td>number of reading sessions</td>
<td>-0.05</td>
<td>-0.007</td>
<td>0.13</td>
<td>0.50**</td>
</tr>
<tr>
<td>number of annotations</td>
<td>0.02</td>
<td>-0.13</td>
<td>-0.07</td>
<td>0.61**</td>
</tr>
<tr>
<td>time relative to class when reading occurs</td>
<td>-0.11</td>
<td>0.13</td>
<td>0.66**</td>
<td>-0.30</td>
</tr>
<tr>
<td>average annotation quality</td>
<td>0.39*</td>
<td>0.36</td>
<td>0.46*</td>
<td>0.14</td>
</tr>
<tr>
<td>ratio of explanations to annotations</td>
<td>0.43*</td>
<td>0.47*</td>
<td>0.26</td>
<td>0.11</td>
</tr>
</tbody>
</table>

*p <0.05  ** p <0.01  *** p <0.001

There is a strong positive relationship between when students read relative to class and their normalized gain scores (0.66 p=0.006). Students who read later (after class) have higher normalized gain scores than students who read earlier (before class). Figure 5.11 shows students’ normalized gain scores plotted against the average time (in hours) relative to class when each student reads. Each point on figure 5.11 represents a student. Generally, the later students read, the higher their normalized gain score are.

Table 5.12 shows standardized coefficients for three separate linear regression
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

Figure 5.11: Normalized gain scores plotted against average time (in hours) when each student reads, relative to class, spring 2013.
Table 5.12: Standardized coefficients for linear regression models predicting post-course CSEM scores. Model 1 controls for pre-course CSEM scores and average annotation quality, Model 2 controls for both pre-course CSEM-scores and the average time (in hours) relative to class when students read. Model 3 controls for pre-course CSEM-scores, average annotation quality and the average time (in hours) relative to class when students read.

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre-course CSEM</td>
<td>0.50***</td>
<td>0.62***</td>
<td>0.59***</td>
</tr>
<tr>
<td>average annotation quality</td>
<td>0.47**</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>average time when students read</td>
<td>0.60***</td>
<td>0.51**</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.42</td>
<td>0.56</td>
<td>0.58</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.81</td>
<td>0.69</td>
<td>0.70</td>
</tr>
</tbody>
</table>

*p < 0.05 ** p < 0.01 *** p < 0.001

models predicting post-course CSEM scores. All three models control for pre-course CSEM scores. Model 1 controls for average annotation quality while model 2 controls for the average time (relative to class) when students read. Model 3 controls for both average annotation quality and the average time when students read.

5.4.2 Discussion

Students who read and annotate more have higher final grades than students who annotate and less less, however the amount time spent reading and annotating is not correlated with any other learning metric (normalized gain, RAA performance, LC performance). This might be due to the fact that the final grade in AP50 is based more on effort than on conceptual understanding. The final grade is mostly a function
of performance on group projects and effort on problem sets and annotations. The other learning metrics (normalized gain, RAA performance and LC performance) are much more indicative of conceptual understanding. In light of this, it is not surprising that the amount of time students spend reading and the number of annotations they write are strong predictors for final grade as, these two metrics are simply proxies for effort.

While the quantity of reading and annotations is not correlated with any of the learning metrics we associate with conceptual understanding (normalized gain, RAA performance and LC performance), the quality of the annotations is predictive of these learning metrics. Students who write high quality annotations have higher normalized gain scores and do better on the individual component of the Readiness Assurance Activities. Students who annotate more with explanations to their peers tend to do better on in-class activities (RAA and Learning Catalytics) than students who annotate mostly by asking questions and making comments. Although these relationships between annotation quality/type and performance on conceptual metrics are not causal, the relationship is still interesting in light of the success of Peer Instruction. Students who engage in high quality discussion online, especially by providing answers to the questions of their peers, perform better on in-class tests of conceptual understanding (RAAs and Learning Catalytics). The fact that these relationships are statistically significant even after controlling for how much physics students know at the beginning of the course (pre-course CSEM scores) is also important to note. It is not the case that only the students who already know physics coming into the course are annotating with high quality explanations and then going
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

on to do well in the RAAs and the Learning Calalytics questions. The relationship between annotation quality and students’ performance on these two activities is independent of how much physics a student knew at the beginning of the semester. While this still does not prove causality, it at least implies that, regardless of their physics background, students who engage in high level discussion on NB, learn more conceptually than students who do not.

The relationship between when students read, on average, relative to the class and their normalized gain scores, is the hardest to explain. The fact that students who read latest (i.e. after the class where the material is covered) also have the highest normalized gain scores is very counter-intuitive. The entire premise of a flipped classroom is that students arrive to class having done the reading and are ready to get the most out of the in-class activities because they are armed with an initial understanding of the chapter that is being dealt with in class. Here it seems as though students who are coming to class the least prepared are gaining the most conceptually.

One possible hypothesis is that students’ reading habits (when they read, for example) is a proxy for some other more powerful predictor of normalized gain. For example, maybe students who know the most physics at the beginning of the course, believe they do not need to read the book before coming to class and these are also the students who have the highest normalized gain. First of all, there is no relationship between how much physics a student knows coming into the classroom and when (relative to class time) they read (correlation =-0.06). Secondly, there is no relationship between pre-course CSEM score and normalized gain so, it is not the
case that the most knowledgable students are reading late and then going on to get higher normalized gains.

Another hypothesis is that students who read late are benefitting from being able to read all the annotations from other students and that, this social intelligence effect is helping students learn.

A third hypothesis is that reading the book after being exposed to activities in class is actually better than reading beforehand. It is possible we are wrong to assume that, to be able to interact productively in-class, students have to read the material ahead of time. Maybe reading the material after engaging in class allows students to read more productively, through a lens of knowing what is important because it was emphasized during the in-class activities.

To shed some light on the puzzling relationship between reading late and having high normalized gain scores that we noticed in spring 2013, starting in fall 2013, we started seeding some of the sections with annotations from the previous year. The manner in which we did this has already been described in the methods section. The reason we did it was, at least in part, to test the social intelligence hypothesis. We wanted to see if students in seeded sections had higher normalized gain scores than students in unseeded sections. If this were the case, it would provide insight into the relationship between reading later and having a high normalized gain score by supporting the social intelligence effect. That is, if student in seeded sections had higher normalized gains than students in unseeded sections, it would suggest that it is not so much when students are reading that affects how much they learn but rather, whether the forum is populated with other students’ annotations.
5.5 What effect does seeding the forum have on student participation in the forum?

5.5.1 Results

We find that there is no significant difference in the CSEM gain between students in seeded versus unseeded sections. Figure [5.12] shows the average normalized gain scores of students in the four sections in spring 2014 (two of which were heavily seeded with annotations from the previous year and the other two of which were unseeded). There is no significant difference in the average normalized gain of students in seeded versus unseeded sections even if we collapse the two seeded and two unseeded sections. The average normalized gain of students in unseeded sections is 0.29 compared to 0.21 in seeded sections (p=0.23).

We find five other interesting results from our experiments with seeding sections. The first is that seeded threads stimulate an above-average amount of discussion, compared to unseeded threads. We find a statistically significant difference in the average thread length of seeded threads receive compared to unseeded (i.e. the usual) threads. Unseeded threads are, on average 0.46 replies, while seeded threads receive an average of 1.16 replies (p<0.0001).

Second, we find that seeded threads demonstrate an above average amount of ‘generative’ discussion. Figure [5.14] shows the fraction of threads that fall into each of the five thread types (from table [5.6]) for unseeded versus seeded threads. We used an ANOVA analysis of variance to determine that the difference between groups is statistically significant at the p<0.05 level. Especially noteworthy is the large fraction
Figure 5.12: Average CSEM gain in seeded versus unseeded sections, spring 2014
Figure 5.13: Average number of replies in unseeded (white) versus seeded (grey) threads
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

Figure 5.14: Different types of thread types between seeded and unseeded threads

of elaborative generative discussions that emerge in the seeded threads compared to the unseeded threads.

Third, we find that, on average the quality of the annotations in the seeded sections and threads exceeds the quality of annotations in the unseeded sections and threads. Figure 5.15 shows that the quality of annotations in seeded sections (1.58) exceeds the quality of the annotations in the unseeded sections (1.56), which is statistically significant at the p<0.05 level. This figure also shows that the quality of the annotations in the seeded threads (1.71) is significantly higher than the quality
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

Figure 5.15: Average quality of annotations in unseeded versus seeded sections (left) and in unseeded versus seeded threads (right)

Fourth we find that the quality of annotations is highest in sections seeded with the 10 longest threads (seeding condition 1 from table 5.4). Figure 5.16 shows the average annotation quality in sections with each of the four seeding conditions outlined in table 5.4. The quality of annotations in the sections that were seeded with the 10 longest threads was statistically significantly higher than sections that were not seeded at all (p<0.05) and sections that were seeded with the 10 highest quality annotations (p<0.01). Based on this, we seeded with the longest threads in the third
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

Finally we find that annotations in seeded section are made significantly earlier than annotations in unseeded sections. In fall 2013/spring 2014 annotations, in seeded sections, were made 21 hours before class whereas annotations in unseeded sections were made 11 hours before class (p<0.001).

Figure 5.16: Average annotation quality in sections with different seeding conditions

semester of our study (spring 2014).
5.5.2 Discussion

In fall 2013, we found that students who read late, relative to the class time, have higher normalized gain scores than students who read early (see figure 5.11). We proposed a few different hypotheses to explain this relationship, one of which is there is a social intelligence effect. To test this, in spring 2014, we split the class into four sections and seeded each of the chapters in two of the sections, with the 20 longest threads, from the previous year. We can eliminate the social intelligence effect hypothesis, given the fact that students in the seeded sections have normalized gain scores that are not statistically different from students in the unseeded sections. Students who read the book that had been heavily seeded did not have higher normalized gain scores than students who read the unseeded version of the book. In fall 2013 and spring 2014, the relationship between reading late and high normalized gain completely disappeared. There was no difference in normalized gain scores between students who annotated early and students who annotated later. This might be due to the fact that, due to the modified timeliness rubric, in fall 2013 and spring 2014, students read a lot earlier than they did in spring 2013. It is possible that our third hypothesis is the correct one, and that students learn more when they read the textbook after being exposed to the in-class activities. Changing the NB assessment deadlines to incentivize reading before class caused students to read earlier and maybe that is why relationship between reading time and normalized gain scores disappears in fall 2013 and spring 2014.

Despite the fact that students in seeded sections do not have higher normalized gains than students in unseeded sections, there are many other position effects of
seeding. We have provided experimental evidence that it is possible to seed prior-
semester comments into the new semester to stimulate an above-average amount of
discussion, and that this discussion demonstrates an above average amount of ‘gen-
erative interaction’, the interaction type that prior literature \cite{83} has demonstrated
to be of the greatest value for learning. We have also shown that students in seeded
sections annotate and read significantly earlier than in unseeded sections. This inform-
ation is useful to instructors of flipped classrooms who are interested in increasing
the likelihood of their students reading the material before coming to class.

5.6 What effect does varying the sizes of sections
have on student participation in the forum?

5.6.1 Results

We find that the size of the section is positively correlated with the length of
the average thread in that section. We computed the average thread length in each of
the sections as a function of the size of the section. Figure \ref{fig:5.17} shows that the average
thread length increases as the size of the section increases. The correlation between
section size and average thread length is 0.74 (p<0.001).

5.6.2 Discussion

As the size of the section increases, students initiate threads less and instead
add on to existing threads. The correlation between the number of initiated threads
per student and section size is -0.30 (p<0.03) and the correlation between the number
Chapter 5: The role of an online collaborative textbook annotation tool in a flipped, introductory, physics class

Figure 5.17: Average thread length for sections of varying sizes
of replies per student and section size is (0.56 p<0.001).

This finding lends support to the hypothesis that, when there are too many participants in a forum, it gets saturated with annotations and there is nothing left to say. In larger sections, students might be adding comments to existing threads rather than starting their own threads due to this saturation effect. It remains to be determined whether “saturation” forces students to reply to threads instead of initiating their own is a good or a bad thing. As we have argued, conversation has been recognized as having higher impact for learning, which suggests it could be beneficial to force students into conversation. However, if those forced conversations are filled with “me too” statements reflecting the student requirement to comment, then there may be no beneficial dialog.

5.7 Conclusion

Online forums, have large implications for the scalability of a class. An effective online forum could easily replace a traditional, physical section or recitation. A successful, content-based online forum that works well for a class of 90 students could easily be scaled to a MOOC of 100,000. Researching best practices in online forums is essential to making this scalability effective. Through this preliminary look at three semesters worth of data, we have an increased understanding of how students use NB, in the context of a flipped introductory physics classroom and how this use relates to student learning in the class. We have also uncovered some interesting trends regarding the optimal size of a forum. Through seeding discussions with thought-provoking content, we have also gained insight into how students’ online
conversations can be managed and guided to promote learning. Seeding forums with successful threads appears to provoke students to have more constructive (generative discussions). The overall quality of the annotations appears to also be higher in these seeded sections.
Chapter 6

Summary and implications for instruction

The goal of this dissertation is to broaden our understanding of interactive teaching strategies, in the context of the introductory physics classroom at the undergraduate level. We discussed four main projects, each of which investigated a specific aspect of teaching physics interactively. Through a better understanding of the role that student characteristics play in learning, each project indicates specific recommendations for improving interactive teaching methodologies. Here we provide a summary of the findings from each project as well as the implications that each has for instruction.

6.1 Summaries and implications for instruction

In Chapter 1, we study lecture demonstrations in the context of an interactive classroom using Peer Instruction. We investigate the role of students’ predictions and prior
conceptual understanding in their ability to correctly observe and remember physics demonstrations shown to them in class. We measure students’ understanding of the specific concept underlying a lecture demonstration by asking a series of conceptual questions. We also ask students to predict the outcome of the demo. We study the relationships between students’ level of understanding of the concept, their predictions of the demonstration and their ability to observe and remember the demonstration accurately.

We find that students’ initial conceptual understanding of a concept plays a role in correctly observing the demonstration, even when the demonstration is presented in an interactive manner. When asked within a day of seeing a demonstration, students report an incorrect outcome 18% of the time, regardless of whether they predict the outcome correctly or incorrectly. We also find that for a demonstration to lead to conceptual learning, it is not enough to present the demonstration interactively. It is essential that the student observes it correctly in the first place. Students are more likely to observe a demonstration correctly if they a) make a prediction first and b) have some conceptual understanding of the underlying physics beforehand.

These findings support the importance of having students predict the outcome of a demonstration, regardless of whether they predict correctly or not. Furthermore, demonstrations are most effective in promoting learning when students have at least a basic level of conceptual understanding beforehand. In light of these findings, we recommend that demonstrations be integrated into a learning sequence that helps students develop background knowledge first. These findings support the effectiveness of POE strategies such as Interactive Learning Demonstrations [48], which emphasize
both prediction making and conceptual scaffolding of the demonstration.

In Chapters 3 and 4 we analyze response timing and switching patterns to conceptual questions posed over the course of the semester, in introductory physics classes taught using Peer Instruction. We look at how both student response times and switching response patterns relate to student characteristics such as pre-course physics knowledge, science self-efficacy and gender.

For two different student populations, we find that students with greater incoming physics knowledge and higher self-efficacy respond faster, indicating that response times are partly a function of student characteristics. We also find that there is no gender difference in response rate when other student characteristics (self-efficacy and incoming physics knowledge) are controlled for. Students with low self-efficacy also are more likely to switch the right answer to the wrong one. These findings indicate that how quickly students respond and whether they switch their responses to questions during Peer Instruction is influenced by how they perceive their own abilities.

We also find that incorrect answers take more time than correct answers, especially for easy questions. This suggests that incorrect responses result from students not knowing the answer, rather than from strongly held misconceptions. Response switching is also related to the difficulty of the item. As the difficulty of the item increases, so too does the percentage of students who switch their response.

In light of our findings on response timing, we recommend that instructors terminate polls once 80% of the answers are in, because at that point an increasing fraction of students respond by random guessing. Understanding that students switch
to the wrong response more often with difficult questions is informative because it indicates that instructors may need to provide better scaffolding for those questions. Instructors should try to support students through more difficult questions by providing mini-lectures, more pre-class reading materials, or by keeping the polls open longer. Instructors should also consider scaffolding the most difficult CTs by building up to them with a series of less difficult questions. The strong connection between CT switching and self-efficacy has important implications for Peer Instruction. Interventions to increase student self-efficacy at the beginning of the semester might improve students’ experiences during Peer Instruction and help students take better advantage of this teaching strategy.

In chapter 5, we discuss work on the role of NB, an online collaborative textbook annotation tool, in a flipped, project based, physics class. We analyze the relationship between students' level of online engagement and traditional learning metrics to understand the effectiveness of NB in the context of flipped classrooms. We also report the results of experiments conducted to explore ways to steer discussion forums to produce high-quality learning interactions. Through this work, we have an increased understanding of how students use NB, in the context of a flipped introductory physics classroom and how students' level of engagement relates to their learning in the class. We have discovered that, when it comes to online engagement, it is not the quantity of the interactions that contribute to learning but rather, the quality of the interactions that matter. The number of annotations that a student makes or the number of hours spent reading does not correlate with learning metrics. The quality of the annotations that a student makes however, are highly correlated with
performance on interactive in-class activities and conceptual learning in the course overall. We have also uncovered some interesting trends regarding the optimal size of a forum. Through seeding discussions with thought-provoking content, we have also gained insight into how students’ online conversations can be managed and guided to promote learning.

These findings have important implications for incentivizing and managing online reading forums, the use of which is increasing with the recent movement towards flipped classrooms. We have shown that for reading forums to serve as a valuable learning tool, it is essential for them to be managed and assessed carefully by instructors. This includes providing students with clear expectations for both the quality and deadlines of the expectations. Providing regular feedback is also essential in ensuring that students participate actively in the forum. To prevent saturation, instructors should limit the effective size of the forum to around twenty students (through sectioning). Seeding forums with successful threads appears to provoke students to have more constructive (generative discussions). The overall quality of the annotations appears to also be higher in these seeded sections.
A.1 E&M Demonstration Example

A.1.1 Pre-demo Conceptual Questions

A constant potential difference $V$ is applied across two resistors connected in parallel as shown.

The current through the 2 W resistor is 2 A. What is the current through the 4 W resistor?
1. 0 A
2. 1 A
3. 2 A
4. 4 A
5. Need to know the potential difference.
A constant potential difference is applied across two resistors connected in series as shown. The current through the 2 ohm resistor is 2 A. What is the current through the 4 ohm resistor?

1. 0 A
2. 1 A
3. 2 A
4. 4 A
5. Need to know the potential difference.
A.1.2 Predict Question

The light bulbs in the circuit are identical. When the switch is closed

1. both will go out
2. the intensity of light bulb A will increase
3. the intensity of light bulb A will decrease
4. the intensity of light bulb B will increase
5. the intensity of light bulb B will decrease
6. the intensity of both bulbs will remain the same
A.1.3 Post-demo Conceptual Questions

An ammeter A is connected between points a and b in the circuit below, in which the four resistors are identical. The current through the ammeter is:

1. I/2
2. I/4
3. zero
4. need more information
An ammeter A, which measures current, is connected between points a and b in the circuit below. It is noted that the current through the ammeter is zero. Rx is a resistor of unknown value, whereas R, 2R, and 3R are known. What is the value of Rx?

1. 2R
2. 3R
3. 6R
4. insufficient information about the ammeter
5. insufficient information about the source V0
A.1.4 Outcome Question

Recall the demonstration where two identical light bulbs were connected in the circuit depicted below. When the switch was closed, what happened?

1. both bulbs went out
2. the intensity of light bulb A increased
3. the intensity of light bulb A decreased
4. the intensity of light bulb B increased
5. the intensity of light bulb B decreased
6. the intensity of both bulbs remained the same
7. I did not see this demonstration
A.1.5 Explain Question

Using your understanding of the physics principles involved, explain why what you observed during this demonstration took place.

A.2 Mechanics Demonstration Example

A.2.1 Pre-demo Conceptual Questions

While sledding down a steep hill, a boy tilts his head back and spits up in the air at an angle that is perpendicular to the hill. (Neglect air resistance and friction) Which of the following statement(s) are true about the motion of the spit and the boy. (Select any that apply).

1. the spit will narrowly miss the boy, landing just behind him
2. the spit will land directly on the boy
3. the spit will land in front of the boy
4. the boy is accelerating down the incline but the spit (once it leaves the boy’s mouth) is moving at a constant speed (straight up and straight down)
5. the boy’s speed is constant but his spit has acceleration due to gravity
6. the boy is accelerating down the incline and his spit (once it leaves the boy’s mouth) is accelerating at the same rate in the direction of the incline
A ball is dropped from the top of the mast of a sailboat floating quickly down a river. Where will the ball land relative to the mast?

1. in front of the mast
2. at the base of the mast
3. behind the mast
4. it depends on how quickly the boat is traveling
A.2.2 Predict Question

A ball launcher that sends balls vertically (perpendicular to the plane) is sitting on a cart. The cart is released from the top of an inclined plane when the ball is projected. Where will the ball land relative to the cart?

1. in front of the cart
2. on the cart (at the base of the launcher)
3. behind the cart
4. it is impossible to say
A.2.3 Post-demo Conceptual Questions

A flare is dropped from an airplane flying at uniform velocity (constant speed in a straight line). Neglecting air resistance, the flare will:

1. quickly lag behind the plane
2. remain vertically under the plane
3. move ahead of the plane
4. it depends how fast the plane is flying

A cart with a ball-launcher starts from rest at the bottom of an inclined plane and is pushed so that it travels up the inclined plane. Halfway up the plane, the cart shoots the ball upwards (perpendicular to the plane). The cart continues to the top of the track at which point it changes direction and rolls back down the track. Where would the ball land relative to the cart?
1. the ball will land behind the cart as the cart is still moving up the incline
2. the ball will land behind the cart as the cart is moving down the incline
3. the ball will land in the cart as the cart is moving down the incline
4. it depends on how high the ball is projected as to where it lands relative to the cart
A.2.4 Outcome Question

Recall the demonstration where the cart moved down an inclined plane and projected a ball upwards (perpendicular to the track): Where did the ball land relative to the cart?

1. in front of the cart
2. on the cart (at the base of the launcher)
3. behind the cart
4. I did not see this demonstration

A.2.5 Explain Question

Using your understanding of the physics principles involved, explain why what you observed during this demonstration took place.
Appendix B

Appendix 2

Right to Right (normalized) = \frac{\text{Right} - \text{Right}}{\text{Right} - \text{Round1}}

Right to Wrong (normalized) = \frac{\text{Right} - \text{Wrong}}{\text{Right} - \text{Round1}}

Wrong to Right (normalized) = \frac{\text{Wrong} - \text{Right}}{\text{Wrong} - \text{Round1}}

Wrong to Wrong - Same (normalized) = \frac{\text{Wrong} - \text{Wrong}\text{(same)}}{\text{Wrong} - \text{Round1}}

Wrong to Wrong - Different (normalized) = \frac{\text{Wrong} - \text{Wrong}\text{(different)}}{\text{Wrong} - \text{Round1}}
Appendix C

Appendix 3

C.0.6 Self-Efficacy Survey

Rank your level of agreement with each of the following statements using the following scale:

1: strongly disagree
2: disagree
3: neutral
4: agree
5: strongly agree

1) I enjoy learning about science
2) I enjoy learning about physics
3) I often do well in science courses
4) I often do well in non-science courses
5) I identify with students who do well on exams and quizzes in science courses
6) I expect to receive an A- or higher in this course
7) I am confident I can do the work required for this course
8) Doing laboratory experiments and write-ups comes easy to me
9) I am often able to help my classmates with physics in the laboratory or in section
10) I usually don’t worry about my ability to solve physics problems
11) When I come across a tough physics problem, I work at it until I solve it
12) I get a sinking feeling when I think of trying to tackle difficult physics problems
13) I like hearing about questions that other students have about the reading
14) I am usually confident of my answers to the EARS questions before I talk to a neighbor
15) I am usually confident that I can convince my neighbor of my answer to EARS questions
16) I know how to explain my answers to EARS questions in a way that helps others understand my answer
17) My peers know how to explain their answers to EARS questions in a way that helps me understand their answer
18) Listening to my neighbors talk about their answers increases my confidence when responding to the same EARS question a second time
19) Practicing answering EARS questions in class makes it easier for me to do physics problems at home
20) I can communicate science effectively
21) I can communicate physics effectively
C.0.7 Annotation Example

Figure C.1: Annotation examples with explanation of the Year 2 quality rubric
Bibliography

[1] President’s Council of Advisors on Science and Technology, 2012.


Bibliography


[70] M. Sorcinelli, A.W. Chickering, and Z.F. Gamson, Applying the Seven Principles
for Good Practice in Undergraduate Education, New Directions in Teaching and Learning (Jossey-Bass, ADDRESS, YEAR).


[79] Boulder University of Colorado, PhET Interactive Simulations.

