TOOLS FOR UNDERPREPARED STUDENTS IN ENGINEERING PHYSICS WITH A FOCUS ON ONLINE MASTERY LEARNING EXERCISES

BY

BRIANNE GUTMANN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Gary Gladding, Chair
Professor Tim Stelzer, Director of Research
Professor George Gollin
Professor Cynthia D'Angelo

Abstract

For students entering college with very little or poor physics background, the transition into fast-paced calculus based physics courses at a large university can be demanding; this thesis documents tools used in a preparatory course which is intended to ease that transition by teaching fundamental skills and problem solving. Within the preparatory course, we implemented online mastery-style homework, introduced frequent testing with retakes, and applied a values affirmation intervention intended to quell stereotype threat. In the initial implementation of our mastery-style exercises, which were successful in two clinical trials, mastery was less successful in the course due to high levels of student frustration. Adjustments to the delivery and content with student affect in mind were able to mitigate frustration and increase students' productive behaviors and performance. On mastery activities which were particularly difficult or calculation-heavy, increasing the amount of scaffolding and breaking up content into smaller, more focused pieces were strategies that helped students master content. Frequent testing improved midterm scores, but the effects were diminishing with the number of tests. The values affirmation activity, which was given to the preparatory course and a more mature physics course, had little effect on students' performance despite its success documented elsewhere. Reflection on these tools' success has highlighted the importance of considering student affect and the sensitivity of their effectiveness to students' sense of agency and to the tools' specific implementation.

To my parents, my first and forever heroes.

"Anyone can slay a dragon, he told me, but try waking up every morning & loving the world all over again. That's what takes a real hero."

Acknowledgements

This work is the culmination of many years of support and guidance from colleagues, friends, and family, for whom I have sincere love and gratitude. I would like to acknowledge these people and honor the gifts they have extended to me.

Foremost, the work in this thesis would not be possible without my research advisor, Tim Stelzer. His genuine and holistic approach to advising has been a comfort and have helped me grow as a person and as a researcher. He has been exceedingly kind and a strong ally to my success, and ultimately my well-being. I am grateful for his sincerity and openness, his sneaky humor, and consistent and flexible support. He has always understood my interest in community building and allowed me the space and time to pursue the things that give me energy, while also ensuring that I don't lose sight of my research goals. In research, he has taught me to be critical, to assert my ideas, and to tell a story. I would not be where I am without his mentorship and I am in awe of the care he put into cultivating our relationship and my place in the research community.

Within our Physics Education Research group, I am also grateful to Gary Gladding, who guided my initial entrance into mastery-style exercises; all our endless days and nights creating content and coding problems made this project a reality. Likewise, the contributions of Morten Lundsgaard, the instructor of Physics 100 and our teammate, have been instrumental to its success. The graduate students in our research group, both past and present, have made my time enjoyable and productive by being excellent friends and colleagues. Thank you to Noah Schroeder, Zhongzhou Chen, Witat Fakcharoenphol, Sara Rose, Katie Ansell, Jason Morphew, Bill Evans, Sam Engblom, Devyn Shafer, Muxin Zhang, and Gabe Ehrlich. In particular, I'd like to thank Noah, who worked with me on early mastery projects and helped me understand graduate research in my first years. Without her database magic, I would also not have been able to complete this project without Rebecca Wiltfong, whose skills and warm friendship have led to many giant excel spreadsheets and afternoon bubble teas. Financial support for my work and this thesis has been funded in part by the National Science Foundation, NSF DUE 16-08002.

My colleagues in The Access Network have also been excellent friends and role models, demonstrating the potential for equity work in education and guiding my thinking while providing me opportunities to act on that commitment. Thank you to Chandra Turpen, Joel Corbo, Gina Quan, Dimitri Dounas-Frazer, Ben Pollard, Corey Ptak, Scott Franklin, and Angie Little for your trust, encouragement, and connections to like-minded researchers in physics education. Thank you also to our graduate head at the University of Illinois, Lance Cooper, for encouragement and full endorsement of equity projects in our local department. Without his enthusiastic support and open mind, and without logistical support from Wendy Wimmer, they

could not have been nearly as successful.

To reach this point in my life, however, some of my greatest thanks go to my parents, Glen and Marilyn Gutmann, who believed in and encouraged me with such ferocity that I was unable to believe otherwise myself. They have taught me to temper ambition with humility and self-care, but that I could achieve whatever I set my mind to and more. Most importantly, they taught me to enjoy life's small pleasures and face each day with optimism and integrity. I love you both so much and am grateful to the example you set: to live each moment honestly and kindly, to be gentle and generous with all people, and also to stand up to injustice and be firm in that. I am so proud to be your daughter.

In the vein of honoring simple joys, I would like to thank my friends who have contributed so much joy to my life in the context of small moments. Thank you to Cheryl and Pete, some of my oldest and closest friends. We are each on our own adventures, but you will always be the friends I return home to, who have seen the most and been there from the beginning. While graduate school can sometimes feel never-ending, my time here has been measured by countless wonderful memories with local friends: learning from Courtney how to slosh water in our bellies; knitting and gossip with Hattie, Alisha, Hahna, and Elizabeth; lots of Jeopardy, garlic bread, and cat birthday parties with Hong, as well as our cruising bike rides, wood-carving, and treetop picnics with Apoorv and Jasmine. I am grateful for Revenge bingo and optimum couponing with John, and endless feasts and bonfires with Phil, Megan, Srinidhi, and Mohammed. In the last months of this dissertation, I am especially thankful to Phil, Megan, and Gloria for snacks, coffee, and companionship as we camped out to work.

To those who have worked beside me in organizing, thank you for showing me the power of people working together and validating that work. To Gloria, Will, and Angela, thank you for including me in the creation of Illinois GPS, the program which acted as a catalyst for the rest of my organizing and for which I have utmost love. Illinois GPS introduced me to my mentee Jake, who inspires me to keep pushing and to sometimes indulge in sandwiches covered in mozzerella sticks. Thank you to Karmela, for conversely feeding me vegan food and kombucha, and for maintaining and bringing new energy to organizing in the physics department. To my fellow GEO organizers, being part of our union has made me proud and better. Especially to my fellow physics stewards, Roshni and Ryan, and our union staff, Natalie and Andrea, thank you. It has been stressful at times, but ultimately an experience that was more rewarding than I could have anticipated. I am also grateful to the LGBT folks in physics and math for embracing our collective and making me feel stronger and safer in my identity. There are too many people to thank properly from these last years, but I am full of love for the people who shared their time with me.

Finally, the quiet MVP of my dissertation is my partner and best friend, Jordan. Thank you for loving me, feeding me, and keeping me sane in these last years, and especially these last months. Although you don't ask for recognition often, I see and appreciate everything that you do and the care that goes into it. My jordie, my doggo, my teammate, my zealous protector and love of my life, I cannot wait to start our next adventure together. A love laser to you: (>ˆ_ˆ)> <3 <3 <3 <3 <(ˆoˆ)>

Table of Contents

Chapter 1

Introduction

At a large university, an incoming population of students is inevitably going to have a diverse background of experience and preparation. This diversity creates an incoming class with a spectrum of familiarity with concepts, abilities in problem solving, and fundamental skills. Despite this, most students in engineering or physics are set along the same introductory physics sequence, treated to the same material regardless of the depth of their preparation in high school. To counter this problem, the physics department at the University of Illinois at Urbana-Champaign created a preparatory course called Physics 100: *Thinking About Physics*, recommended to incoming freshmen who perform below a defined threshold on a physics diagnostic test prior to arrival, or who simply feel underprepared for the calculus based physics sequence. Within this course, effort has been put into helping students who are at risk to fail. While creating course materials, one hopes for an approach that will cater to as many students as possible, while also giving support to the most vulnerable students. This dissertation documents and evaluates course changes in Physics 100 which intended to support its students through the transition into college and on to the calculus-based physics sequence.

1.1 Population and Retention

The population of students taking Physics 100 are primarily first-term freshmen, who are encouraged to take the course based on their performance on a diagnostic physics exam which they take online during the summer. The diagnostic test has 13 problems and 3 survey questions about students' previous experience with physics. The problems include some calculations but focus on conceptual understanding of mechanics and ability to write and manipulate symbolic equations. The full content of the diagnostic exam is provided in Appendix A. Students who correctly answer fewer than 8 questions are placed in the course by their academic advisers, and students who correctly answer 8-9 questions are recommended to take Physics 100; the course is not technically mandatory, but most students who are advised to take the course do. Students who score above the thresholds are still welcome to take the preparatory course if they are concerned about their preparedness, and will be advised to take it if they do not have the appropriate math prerequisites.

Looking at the population from the diagnostic test in 2017, the majority of students who enrolled in Physics

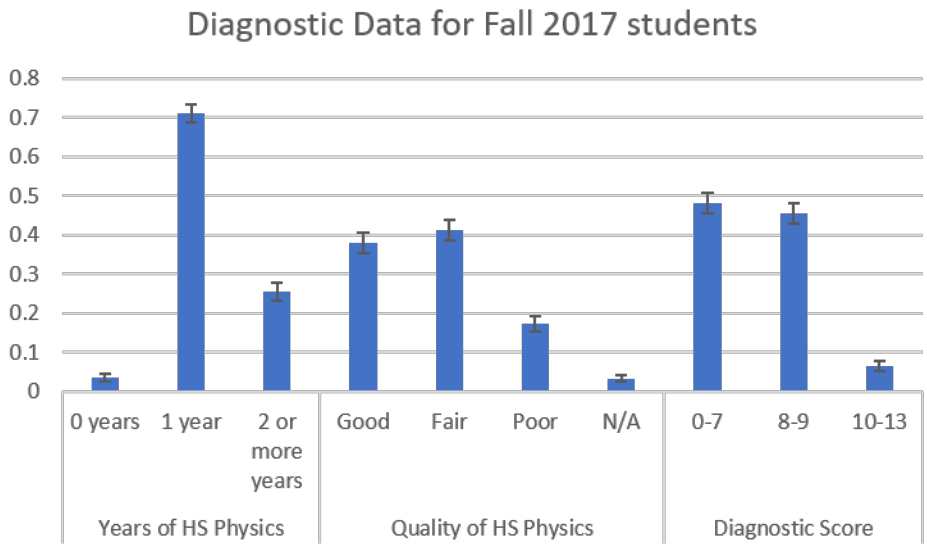


Figure 1.1: Responses to the survey questions and diagnostic score for students entering Physics 100 in Fall 2017.

100 in Fall 2017 had one year of physics prior to entering college, and most rated that course as being "Good" or "Fair" quality, though 17% rated their previous courses as "Poor." The most likely reason students signed up for the course were their scores on the diagnostic; about equal numbers of students scored within the 0-7 range and the 8-9 range, both of which would prompt their academic adviser to suggest the course, and a small fraction scored 10 or more points. Students' distribution of responses to the survey questions and overall scores are shown in Figure 1.1. Students who do not take Physics 100 begin their engineering physics sequence with Physics 211: *Introduction to Classical Mechanics*, the target course for the preparation in Physics 100. The typical student takes Physics 211 in the spring semester of their first year, so taking Physics 100 in the fall of their first year should not set students back in their coursework. The calculus based physics sequence is a requirement for most majors in the engineering college, not only physics majors.

Using data from enrollment from Fall 2011 until Fall 2018, the demographics of students in Physics 100 can be compared to students in Physics 211, which is a proxy for the engineering college at large. Considering enrollments over these eight years, self-reported race and gender demographics of enrolled students are summarized in Table 1.1. Students' options for self-reported demographics are limited to options provided by the University of Illinois; students' genders are restricted to binary options and demographic information for International students does not allow for more specific designation than "International." As students are choosing their identities themselves, it is also unclear whether international students are more likely to choose their racial identity or their identity as an international student as their demographic. This data also only includes students who completed each course.

It is important to note that students in Physics 100 are often later in Physics 211 as well; these demographics do not represent distinct sets of people. About 78% of students in Physics 100 enrolled in Physics 211, and this portion of students makes up about 18% of Physics 211 over the 8 years. These percentages are representative of the data, but not necessarily behavior; the data set does not include Physics 211 students' behavior prior to Fall 2011 or Physics 100 students' behavior after Fall 2018, so there may be students who

Physics 100	M	F		**Physics 211**	M	F	
International	2.8%	1.1%	**3.9%**	International	14.0%	3.3%	**17.3%**
White	38.1%	18.4%	**56.5%**	White	34.0%	9.8%	**43.8%**
Asian	11.1%	7.2%	**18.2%**	Asian	19.6%	6.7%	**26.3%**
Hispanic	8.9%	3.6%	**12.4%**	Hispanic	5.2%	1.5%	**6.7%**
Black/African American	2.9%	1.3%	**4.2%**	Black/African American	1.4%	0.4%	**1.7%**
Multi-Race	2.3%	1.6%	**3.9%**	Multi-Race	2.7%	0.9%	**3.6%**
NHPI	0.0%	0.0%	**0.1%**	NHPI	0.1%	0.0%	**0.1%**
AIAN	0.0%	0.0%	**0.0%**	AIAN	0.0%	0.0%	**0.1%**
Unknown	0.5%	0.2%	**0.7%**	Unknown	0.3%	0.1%	**0.5%**
	66.6%	**33.4%**			**77.4%**	**22.6%**	

Table 1.1: Percentage of self-identified demographics for students summed over 2011 to 2018. The percentages are out of 2,298 enrollments for Physics 100 and 10,240 enrollments for Physics 211 over the 8 years. Students who took both Physics 100 and Physics 211 appear in both sets. The number of students who are in both data sets is 1,812, which is about 78% of Physics 100 and 18% of Physics 211. (NHPI stands for Native Hawaiian and Pacific Islander, and AIAN stands for American Indian and Alaska Native.)

would be in both courses if the data spanned more semesters. By removing those students, the rest of the students could be categorized as those who took only Physics 100, both Physics 100 and 211, and only Physics 211; differences in demographic makeup of these groups can be seen in Figure 1.2.

Recognizing that only a subset of students from Physics 100 enroll in Physics 211, student retention as measured by success in Physics 211 shows that there is still a significant portion of students who are not able to pass the target course. Figure 1.3 shows the fraction of students in 100 who receive each final grade in 211, in comparison to students who enter Physics 211 directly. About 12% of students in Physics 100 did not enroll in the target course, and 15% did enroll but did not pass, compared to 6% of failing students who enter 211 directly. Of course, there is self-selection between these two groups; students who opt to take Physics 100 have identified themselves as less practiced, whereas students who enter Physics 211 are confident that they do not need extra preparation. Thus, it makes sense that there are more A's and fewer failing grades for students who entered 211 directly. As mentioned at the beginning of this analysis, the data above also includes only students who completed the course, which minimizes the effects of losing students through dropping in each of the courses and loses information for those students.

A more detailed analysis of students who took Physics 100 in Fall 2015 - Fall 2017 looks at the steps along the way where we may lose students, including those who enrolled but did not complete the course. The detailed progression of students enrolling, finishing, and passing each course is shown in Table 1.2 and plotted in Figure 1.4. The largest losses of students occur between enrolling and finishing Physics 100, and between finishing Physics 100 and enrolling in Physics 211. It makes sense that students may drop courses early if they are still finalizing their schedule, so the condition for enrollment for both courses was completing at least 3 homework assignments to account for natural dropping. It is at least encouraging that most who finish Physics 100 pass the course, but because 100 exists primarily as a preparatory course for the calculus based physics sequence (which begins with 211), it is concerning to lose so many students between the two courses.

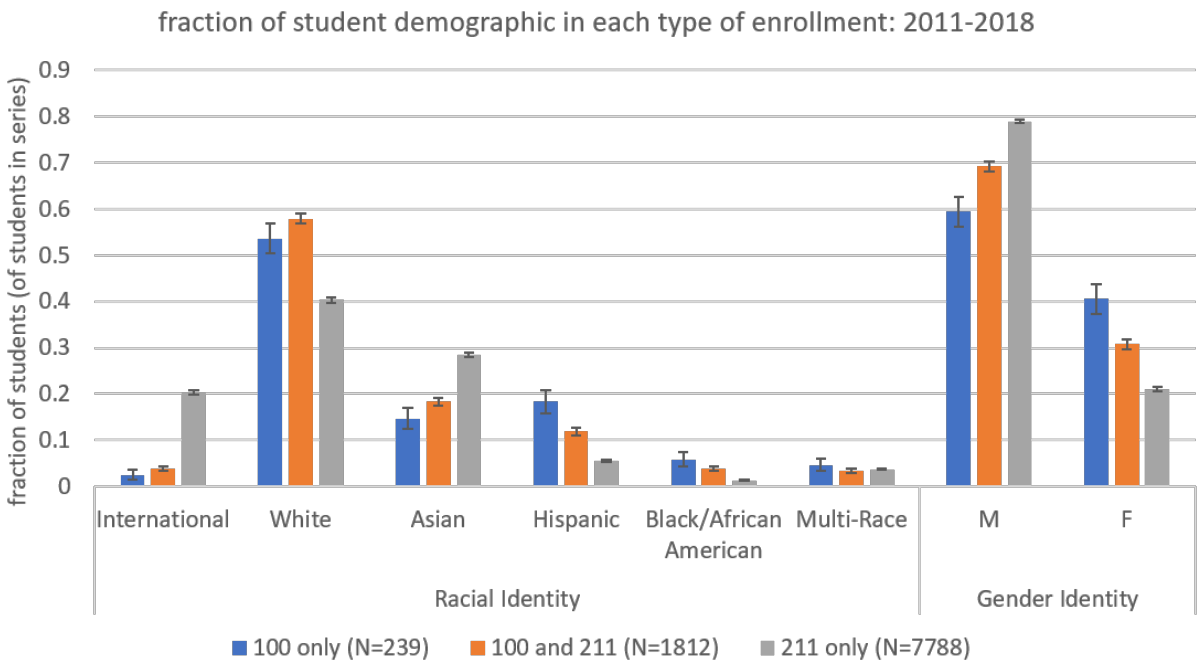


Figure 1.2: Fraction of students self-identified with racial and gender identities within each type of enrollment. The fractions are of the number of students in each category (shown in parentheses in the legend), not fractions of the same pool of students. This data excludes Physics 211 students from Fall 2011 (as there is no information about their Physics 100 enrollment) and Physics 100 students from Fall 2018 (as there is no information about their Physics 211 enrollment). NHPI and AIAN categories are not included in the plot because they cannot be seen on this scale, but their values follow. NHPI students made up 0% of students who took Physics 100 and took both courses, and represented 0.10 $\pm$ 0.04 % of students who took only 211. AIAN students made up 0% of students who only took 100, 0.056 $\pm$ 0.055 % of students who took both Physics 100 and 211, and 0.090 $\pm$ 0.034 % of students who took Physics 211 directly. Students of unknown racial identity make up less than 1% for each cohort.

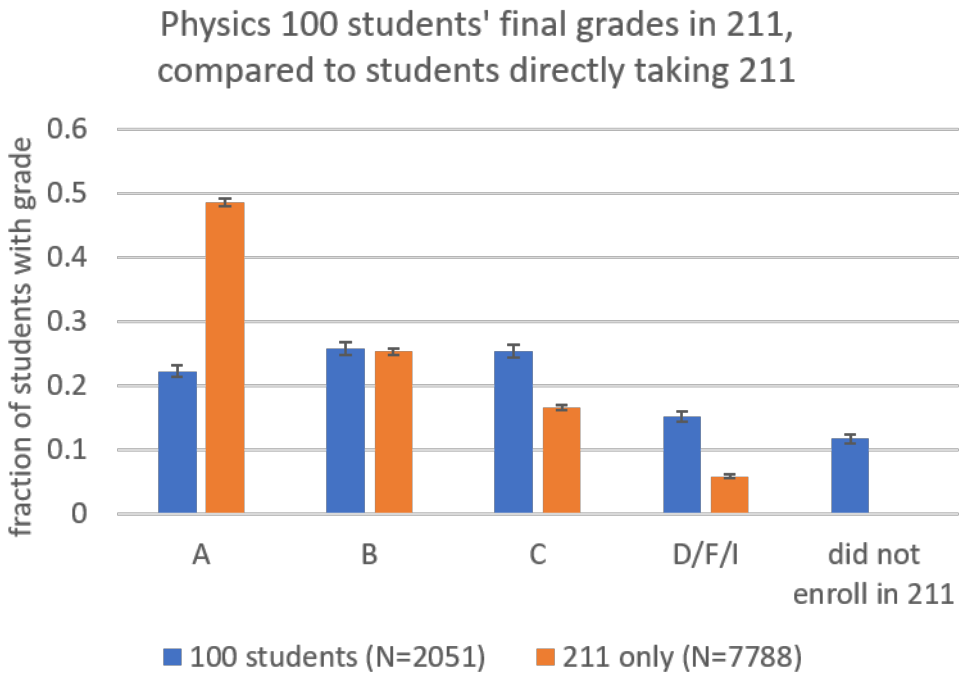


Figure 1.3: Grades in Physics 211 for those who took Physics 100 and those who entered Physics 211 directly. D/F/I includes students who received a D, F, or an Incomplete.

To clarify, these retention rates span years prior to and during the interventions described in this thesis; we did not see substantial changes to retention from our efforts, but low retention rates were a motivation for many of the tools we developed.

In an effort to understand students' motivation for not continuing after Physics 100, a survey was given in 2018 to students which asked them about their plans for Physics 211. The intention of the course is to help any student who would like to pursue engineering; if a student realizes their interests are elsewhere, it is not a bad thing for them to discontinue the sequence, but we would hope that students are not discouraged by the course and are using it as a proxy for the major at large, especially if they have the ability to succeed in 211. Student responses to the survey are in Table 1.3, which showed a higher percentage of students intending to take the course than historically carry on from Physics 100. This may show that students' intentions do not line up with the realities of scheduling, as about 66% of these students ended up taking Physics 211 in the immediately following semester (though some may still take it in a later semester). Only a small fraction of students told us that they were switching majors because of the course difficulty, though some who chose that they preferred a new major admitted in the open-ended comments that their decision was in part due to their preference not to take more physics, and one student commented that their adviser discouraged them from continuing engineering despite their ambition to finish the sequence. Additionally, students in the comments revealed an unlisted reason not to enroll in 211: some intended to take an equivalent course through a community college over the summer (either in their hometown or locally), either for convenience or because they believed the class would be easier elsewhere.

Of those who finish 100 and enroll in Physics 211, most (about 80%) can pass the course, but the fraction of failing students is still larger than the fail rate for the course at large. Again, these are selectively students

semester in 100	**2015**	**2016**	**2017**	**Sum**
enrolled 100	529	437	451	**1417**
finished 100	476	407	394	**1277**
passed 100	459	388	381	**1228**
enrolled in 211	334	306	284	**924**
finished 211	323	290	274	**887**
passed 211	260	233	219	**712**
passed 211 after passing 100	253	227	207	**687**
failed 211 after passing 100	53	53	45	**151**
fraction of 100 (enrolled) who enrolled in 211	0.63	0.70	0.63	**0.65**
fraction of 100 (passed) who enrolled in 211	0.73	0.79	0.75	**0.75**
fraction of 211 (enrolled) who passed 211	0.78	0.76	0.77	**0.77**
fraction of 100 (passed) who passed 211	0.83	0.81	0.82	**0.82**

Table 1.2: Numbers of students who began Physics 100 in Fall 2015 through Fall 2017, and their progression through both Physics 100 and 211. The table also includes fractions of students who enrolled in 211 after enrolling or passing Physics 100, and fractions of students passing 211 after completing 100 and enrolling in 211, or passing 100 and enrolling in 211.

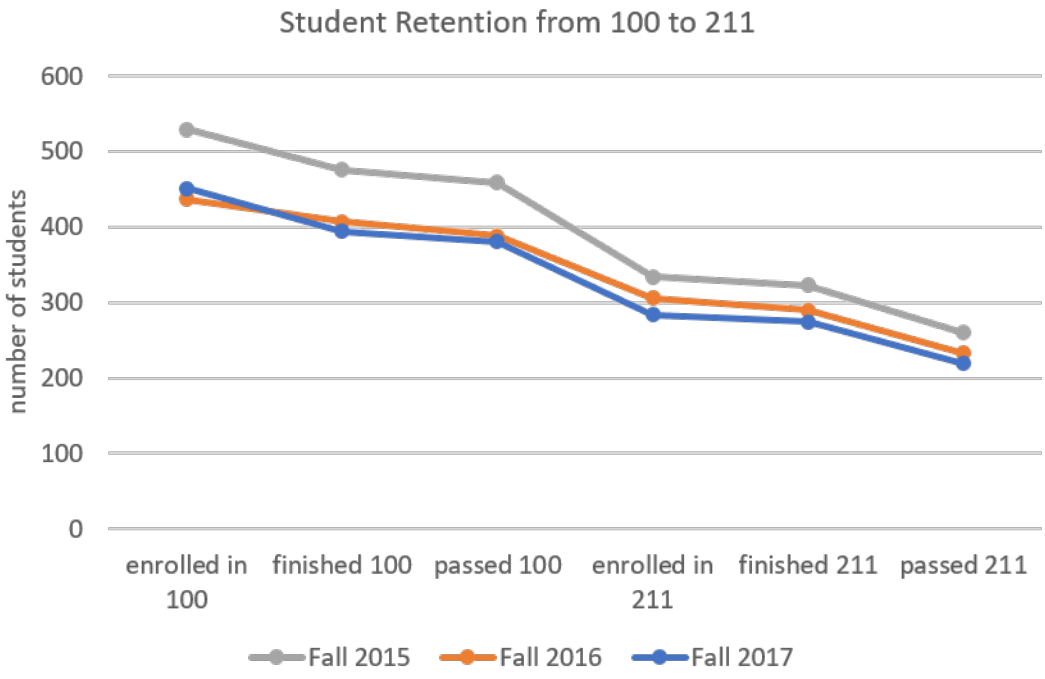


Figure 1.4: Progression of students in 2015-2017, starting in Physics 100 through passing Physics 211.

Are you taking Physics 211 next semester? (N=332)	
Yes! / No, but for scheduling reasons. I plan to take Physics 211 later.	91.5 %
No, I discovered I prefer another major that doesn't need the engineering physics courses.	3.9%
No, I found this course too difficult and have decided to change my major.	0.9%
No, I am going to retake Physics 100 and take 211 after that.	0%
No, but for a reason that is not listed.	3.6%

Table 1.3: Student responses to the question "Are you taking Physics 211 next semester?" from Fall 2018. This survey was given at the end of the course, so does not include students who dropped the course.

who identify as underprepared and are a different population than those who are confident coming into the university, so one should not expect them to be the same. A single course may not be enough support to get them to readiness for Physics 211, but the attrition of students throughout the process was a large motivation for the work presented in this thesis. The major projects intended to support these students included mastery homework, frequent low-stakes testing, and a values affirmation exercise. The mastery homework focuses on adaptable support to help students solidify basic skills, focusing on supplementing prior instruction with necessary skills that may have been missed in high school. Under 40% of students cited their high school physics courses as "Good," so mastery homework intended to identify students' weaknesses and provide support in practice of those skills. Repetitive testing and re-testing is a feature of mastery homework, but was also enacted in bi-weekly quizzes with retakes, intended again to help students identify weaknesses and motivate improvement in fundamental skills. The values affirmation exercise, which consists of two writing activities, had shown success in past for students who were the minority in courses, either racially or by gender, by attending to stereotype threat. As these students are being given a message that they are "behind" immediately upon arriving at the university, the activity had the potential to mitigate anxiety within Physics 100 students, who also have around double the representation of Black and Hispanic students and 50% more representation of women than the engineering college at large.

1.2 Physics 100 Course Structure

Physics 100 has historically been structured (as most of the introductory physics sequence courses at Illinois) into a repeated cycle of activities which introduce students to new content via online prelectures and checkpoints, followed by lecture, homework, discussion (recitation) sections, and a quiz. This process takes place over the course of a week and repeats. In Physics 100, there are 8 weeks of content which span a little over half the semester. The 8 weeks of content cover about a third of the content in the target course, Physics 211. The rest of the time in the semester is dedicated to problem solving strategies and practice of the learned content. A list of students' activities and their timeline within a typical week are shown in Table 5.1.

The online prelectures are multimedia learning modules (narrated animated videos) which introduce students to content before they come to lecture, and can assign students points for watching and answering questions. In multiple studies, these multimedia learning modules have been shown to be more effective than traditional

Table 1.4: Activities and their timing for a typical unit of content in Physics 100. The last 35% of students grades come from their midterm (15%) and final exam (20%).

Day	**Activity**	**Portion of Grade**
Before Friday	Prelecture and Checkpoint	9.5 %
Friday	Lecture	4.5 %
Monday	Office Hours	–
Tuesday, 8am	Homework Due	15 %
Tuesday	Discussion Sections	20 %
Wednesday		
Thursday		
Thursday, 8pm	Online Quiz due	15 %

textbooks, both for conveying new information and incentivizing student engagement before lecture [1–6]. Because students receive points for watching the prelectures and completing corresponding Checkpoint questions, there is more accountability than in the case of a traditional textbook, which most students admit to rarely using [5]. Students' answers to the checkpoint questions also give the instructor a better sense of what students did or did not understand from the lesson and allows lectures to be more tailored towards supporting students' needs rather than teaching all content from scratch. The lecture addresses specific student difficulties and uses an interactive structure; students talk through problems with their peers and use iclickers to answer conceptual questions in real time.

Students' homework is through the platform FlipitPhysics, which allows students to submit answers and receive feedback online. Between lecture and the homework due date, students also have access to office hours, sessions led by Teaching Assistants and the professor over the course of the day. The specifics of the homework delivery for Physics 100 have evolved in the last years; particularly the introduction and adaptations of mastery-style online homework, which began in 2014 which will be discussed more thoroughly in later chapters. After finishing homework, students attend their discussion section, which is a smaller class setting (a maximum of 24 students) led by a graduate Teaching Assistant. These 50 minute sessions focus on working through problems in small groups and give the students an opportunity to attempt more difficult problems with more personalized attention; the discussion section is meant to be the capstone of the cycle where students can clear up any final confusion and solidify their practice. Finally, students take an online quiz, also through FlipitPhysics, intended to measure their understanding of the unit after completing the learning activities. These quizzes are not timed, but students do not receive feedback on correctness until after the deadline. Once the deadline has passed, students have one week to correct their answers for partial credit.

The course also requires students to take a midterm exam, which is scheduled after Week 7 (of 8 content weeks). If students perform well enough on the midterm, they are offered the opportunity to drop the second half of the course and take their current grade. The course is normally 2 credit hours, but students who drop the course after the midterm receive 1 credit hour for the course instead. Typically, a small fraction of students are given this option, and many students do not drop the course even when offered, as the majority of the work takes place in the first half of the course.

Over the years described in this dissertation, there were some changes to the course structure and grading.

The motivation and effects of those changes will be detailed in later chapters, but include the restructuring of the homework into two smaller assignments, introduction of bi-weekly quizzes to replace the online quiz, and the removal of the early-drop option. The following chapters will describe and reflect on the implementation and iterations of mastery style online homework, from clinical trials through full course implementations, and two attempts to address more difficult mastery levels by providing extra scaffolding and breaking content into smaller pieces. Following the mastery-style homework, the dissertation will describe bi-weekly quizzes and retakes, which extended to exam retakes in the following target course, and finally give results of an attempted values affirmation replication.

Chapter 2

Mastery Background

2.1 History

When educator Benjamin Bloom published his proposal, "Learning for Mastery," in 1968 [7], a key assertion that he made was that *any* student can learn, regardless of inherent aptitude and baseline skills, as long as instructional materials and time spent on concepts were adjusted to match each individual's needs [7, 8]. He was inspired by Carroll's definition of aptitude as the amount of time required to attain mastery, with the additional caveat that adjusting time allows any person to reach it [9, 10]. Rather than giving each student the same treatment, learning could be paced so that students who were comfortable could move on while students who were struggling with specific concepts could spend more time and receive feedback and support until they mastered that content.[1]

To implement this idea, Bloom suggested that concepts be broken down into small units; students could be tested on a single unit, then either move on to the next concept if their understanding was sufficient, or be given feedback and correctives, if not. Once a student believes they have understood the concept and completed the correctives, they have the opportunity to re-test, which can allow them to move onto the next unit, or to receive more practice if the retest shows that the skills have not yet been mastered. This continues until mastery is achieved, and then the student is able to move on to the next unit of material, where they repeat the process. Thus, the testing serves a purpose, rather than reinforces the gradient of different students' expected abilities [7].

Although Bloom is generally associated with coining the term *mastery learning*, he was not the only person to advocate this method. There was a small, unsuccessful movement in the 1920s by Washburne [12] and Morrison [13], but it was revived decades later by psychologist B.F. Skinner. Skinner proposed a delivery method which became known as programmed instruction. His theory for learning was based on the idea of operant conditioning, where positive behavior is reinforced by reward and poor behavior is discouraged by a negative consequence. His critique of traditional classrooms was that students don't receive enough feedback, and the feedback is often delayed, further commenting that a lapse between response and reinforcement can

[1] Much of the mastery background in this chapter also appears in "Mastery-style homework exercises in introductory physics courses: Implementation matters," published in 2018 [11].

destroy the effect of conditioning [14]. Skinner's solution was to break the unit into very small pieces and provide feedback often. In his piece, "The science of learning and the art of teaching," Skinner states, "By making each successive step as small as possible, the frequency of reinforcement can be raised to a maximum, while the aversive consequences of being wrong are reduced to a minimum" [15]. Because teachers themselves are incapable of giving constant feedback to every student, he developed learning machines, the first of which was a small prototype with a screen and knobs; an arithmetic problem would be displayed, and only when the correct answer was tuned on the knobs would the next question become available [14–16].

From Skinnerian principles of programmed instruction, Fred Keller developed the Personalized System of Instruction (PSI), which he published in 1968 in an article titled "Good-bye, Teacher...," [17] the same year that Bloom's Learning for Mastery was published. Keller's system was similar to Bloom's, using small competencies and using testing and re-testing, with intervening treatments at different paces for different students. By creating repetitive opportunities for formative testing, students received many iterations of positive or negative feedback, in line with recommendations by Skinner to take advantage of operant conditioning. Because teachers are physically incapable of giving constant feedback to every individual student, Keller suggested the use of proctors or a computer-based delivery method, which marks a difference between the two approaches. Bloom's mastery learning was motivated from an educational perspective, intended for elementary and middle-school classrooms. The corrective interventions were often group-based interactions, and the method relied on students helping each other. Keller's PSI, conversely, was intended for upper level students, on a more individual level, and developed from psychological principles. Keller's recommended correctives were individual activities, such as reading or problem solving, and the PSI has a more flexible pace, like an open-ended online course. In contrast, Bloom's Learning for Mastery (LFM) is student-paced within each unit, but the teacher must also provide some deadlines in order to keep the class moving forward. Thus, Bloom recommended a 90% threshold for mastery, whereas Keller insisted on complete, 100% mastery of a unit to move on. Their grading methods were also different; Bloom suggested assessing students with a final exam, testing the synthesis of all of the individual units into a whole understanding, while Keller believed mastery of the units was synonymous with mastery of the whole, and suggested that mastering each individual unit was enough demonstration of competence [10]. The crux of both these methods, however, was the idea of allowing students to spend as much or as little time necessary to work through explicit objectives via tests and correctives.

Though there are differences, these two methods agree on some fundamental elements for mastery learning. Both Bloom's LFM and Keller's PSI suggest the following, summarized by Block and Burns [10]:

1. Explicitly specify the course objectives that students are expected to master.
2. Break the course into small units, so students are only required to learn few skills at a time.
3. Require mastery by means of testing, corrective interventions, and re-testing, until students master each unit.
4. Students are assessed by their achievement, rather than achievement relative to others. If all students master all the materials, then it is possible for all students to excel in the course.

2.2 Theoretical Framing

The original framework for LFM and PSI, as discussed, was inspired by psychological principles of learning: Carroll's time on task and operant conditioning, specifically. Beyond that, mastery learning takes advantage of other cognitive and educational tools. By testing and adaptively re-testing, the delivery method exemplifies formative assessment, a tool that Bloom cited in 1968 that is now widely understood to improve learning [18]. Black defines "a key characteristic" of formative assessment to be that "the assessment information is used, by both teacher and pupils, to modify their work in order to make it more effective" [19]. This is in contrast to summative assessment, which evaluates student learning without using the information to inform teaching. The reactive nature of the instruction for both students and teachers is a core principle of formative assessment [20–22], which necessitates feedback to occur during the learning process rather than afterwards. Mastery learning inherently provides feedback to students and instructors during its delivery and adjusts learning tasks to respond to the needs of the student; Black and Wiliam cite formative assessment as a key component of effective mastery learning and dedicate an entire chapter to it in their literature review of formative assessment [18].

Further, the act of testing itself has been shown to improve student performance; the testing effect has been documented in many studies, showing that recalling and encoding information are linked cognitively, and that recalling information often enhances retention and performance more than re-study does [23–28]. One explanation for this phenomenon is transfer-appropriate processing [29], an aspect of the levels-of-processing cognitive framework which argues that "retention [is] determined by how well the processing requirements of the test matched those used originally to encode information" [30], meaning that the more similar the procedures used to encode and recall information, the better students are able to retain the information. By learning through testing, students are at an advantage to recall information later during testing. The testing effect has been shown most effective when students are given feedback, while Merrienboer et al have demonstrated the benefit of delivering feedback "precisely when learners need it" [31], a method exploited by mastery learning.

Mastery learning can also easily facilitate scaffolding within learning activities. Because competencies in later units of content often build on previous levels' competencies, learning can be structured to reduce cognitive load by focusing on new skills methodically and ensuring students are comfortable with baseline skills before introducing new challenges which may otherwise be an overload. Considering Vygotsky's zone of proximal development (ZPD), mastery learning should aim to create assessments that are within the zone between what students can do alone, and what they cannot do at all. By giving students tasks that they are able to achieve with help, then providing that help, the students will be learning within their ZPD [32].

Students' progression through mastery further allows a clean mechanism for students to "grow" in skills, which ideally will reinforce a growth mindset for students, a concept coined by Carol Dweck [33]. While some may subscribe to a fixed mindset, the belief that their intelligence and abilities are inherent qualities, Dweck argues that people who see their intelligence as flexible are more likely to respond constructively to challenges with intentional work and practice rather than frustration or defeat [33]. Generally, she found that those with growth mindset tend to have more adaptive achievement behaviors, which she calls "mastery-orientation", while fixed mindsets more often motivate students to avoid challenging tasks to keep themselves safe from failure; she attributes these differences to students' goal orientations [34]. Dweck and colleagues define *learning goals* to be instances where students are seeking to improve their competence, compared to

performance goals where students seek positive judgments and avoid criticism of their competence [34–37]. Ideally, the low stakes of incorrect answers in mastery systems should minimize fear of criticism and give students multiple opportunities to improve within specific skills, which could encourage an overall growth mindset and learning goal orientation.

2.3 Empirical Evidence

Beyond the reasons mastery learning should be a useful teaching tool, these principles have been applied in many variations and significant research has looked at the effectiveness of the mastery method in practice. Kulik and Kulik, with various collaborators, have done extensive meta-analyses of studies that identify as Keller's Personalized System of Instruction and Bloom's Learning for Mastery, assessing achievement, retention, attitudes, and types of learning [38–41]. Their most recent, published in 1990 with Bangert-Drowns [38], re-evaluated and synthesized one of their earlier works [41] with other meta-analyses done by Guskey and Gates [42] and Slavin [43], evaluating a total of 108 papers. Of these, 72 were college-level PSI studies, 19 were college-level LFM studies, and 15 were LFM studies in secondary school, with a focus on higher-level courses. Only 2 were elementary-level LFM studies. Because there was a large selection of studies, Kulik et al were able to assess mastery learning across many variables, mentioned earlier as achievement (immediate and retained), affect and attitude towards the course and material, and its effectiveness for different types of learning (such as algorithmic skills versus problem-solving).

The most immediate question when assessing learning could arguably be students' achievement; do they learn the material? Most studies cited student performance as final examination scores, and this was the measure that Kulik et al's meta-analysis used. Of the 108 studies, 103 reported exams scores, with 96 studies citing positive effects for mastery-taught students compared to traditionally-taught students, 67 of which were deemed statistically significant. While seven of the studies showed negative effects due to mastery, none of them were significant. The average effect size over all studies was 0.52 standard deviations, and the distribution of effect size is shown in Figure 2.1. Though many of studies included did not follow up to test retention after the end of the course, all 11 studies that did measured an effect size for achievement on the follow-up exam, their average effect size 0.71, with delays on the order of many weeks; some studies have shown mastery learning methods to enable retention of material [38].

Additionally, a review by Block and Burns [10] cites mastery's ability to address education debt[2] by helping both low-performing and high-performing students. Returning to Bloom's assertion that aptitude is not the same as achievement [7], mastery learning has been found to minimize learning differences between students, especially when they are cognitively different [10], but also as a tool when the gaps are due to social disadvantages [10, 45]. Kulik et al's analysis found across thirteen studies that mastery-style has a higher effect size for lower-performing students than higher-performing students, but not significantly [38]. Struggling students have more space to improve, but mastery is effective at all levels.

Not only does mastery seem to help students achieve and retain competence in a subject, but it has also been demonstrated to improve their learning attitudes. Sixteen out of eighteen studies examined in the meta-analysis with affect assessments showed positive attitudes towards the mastery method, and twelve of

[2]Education debt was a phrase coined by Gloria Ladson-Billings as a re-framing of the more common term, achievement gap [44].

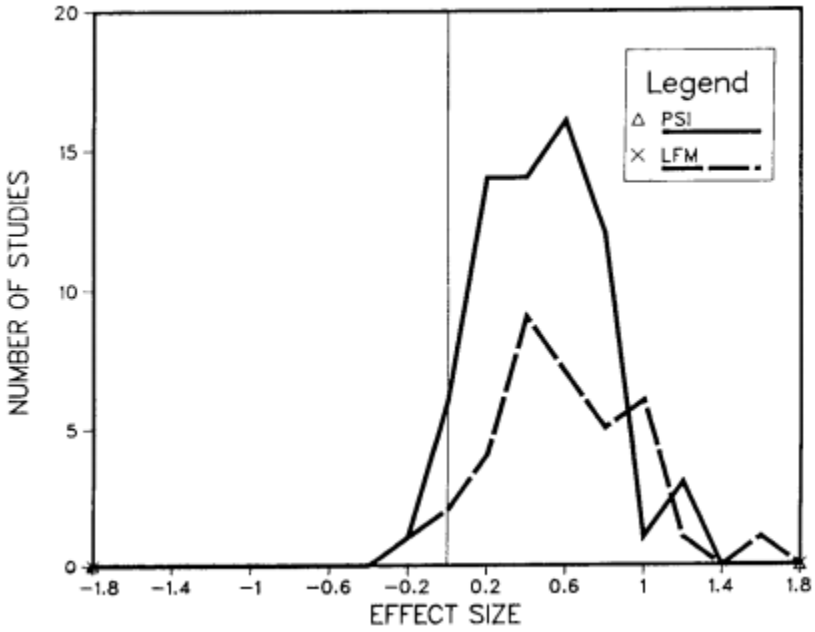


Figure 2.1: Histogram of effect sizes for LFM and PSI studies in Kulik et al's analysis [38]

fourteen reported a more positive student attitude towards the academic subjects themselves [38]. Block and Burns' review cites eighteen separate studies which also reported higher affective responses from students in "interest in and attitudes towards the subject matter learned, self-concept (academic and more general), academic self-confidence, attitudes towards cooperative learning, and attitudes toward instruction" [10]. They also bring up, however, that increased test anxiety may hurt students in a system which relies heavily on testing, but ideally repetition can create familiarity and reduce this problem.

An additional criticism of mastery learning often raised is the belief that the style is only suitable for rigid, algorithmic skills [10, 46]. While repetitive practice may seem well-suited to mastering basic skills, much research has shown that it is also effective at teaching critical thinking. Block and Burns' review uses mastery's tendency to improve performance on essay questions (above the effects for fact-based multiple choice) as an indication that mastery can teach "comprehension, application, analysis, synthesis, and evaluation skills" [10]. Kulik et al found that the largest effect sizes in their analysis were social science courses, rather than math and science classes, and this could be an indication of their conclusion that mastery is actually more suited to higher-level thinking [38]. They attribute this finding to social science's emphasis on analysis rather than basic skills, but Block and Burns further discuss a mastery study in a math classroom where conceptual ideas were excluded for favor of grinding out problems, and mastery students had a diminished positive effect compared to traditional concept-inclusive math courses [10]. In studies that specifically targeted types of learning, there are many which argue that mastery learning is actually better suited for problem solving and higher-order thinking than fact recall [10, 14, 47].

When considering an implementation of a mastery-style physics course, Kulik et al additionally provided which relevant variables created larger effect sizes. Studies which had the biggest positive effects for mastery-taught students shared a high mastery requirement (for example, 90% or 100% performance yielding mastery), locally developed tests (as opposed to national standardized tests), teacher-paced units (instead of

entirely free of time-restraints), and effective correctives [38]. The last point is suggested indirectly by the observation that studies with control groups which did not receive supplemental correctives had a larger effect than studies where both groups were given feedback, which simply reinforces the importance of choosing corrective feedback carefully. One last element to consider is the time required to run each method. While mastery style classes tend to take more time, it has been shown that the proportional time disparity between traditional and mastery delivery methods diminishes as the length of the course increases; the difference is most stark when using mastery for a short time, but is reduced if used for an entire semester [10]. The takeaway of mastery research, as it stands, is that mastery learning can be very effective for maximizing achievement, retention, positive student attitudes, and higher-order thinking skills with only a cost of time.

2.4 Mastery in Physics

Though much of the research using the LFM and PSI has been directed at a variety of courses outside of STEM, there are positive results in mathematics courses, as well as a movement in the 1970s when several universities implemented self-paced introductory physics courses which were modeled after Keller's PSI [48–53]. Although the movement was enthusiastic and college physics seems well suited for mastery, the use of these courses did not become widespread. There have been successful online homework systems with immediate feedback within physics [54–56], but these did not use the mastery delivery system, although they took advantage of some similar learning principles related to online systems with frequent feedback [57–59]. These include the commonly used Mastering Physics system, which is also online with interactive feedback but is not mastery learning as described by Bloom and Keller, despite its similar name [60]. Most of these systems differ from mastery by not providing variation for repeated attempts and correctives. Recently, Heckler and Mikula have implemented an intervention using mastery learning and delivery in the physics domain, but their work focuses on essential skills, specifically vector math as is applied to physics, rather than holistic content for a physics course [61, 62]. They describe their work as similar to the online intelligent tutor ALEKS, which also uses mastery for math skills [63].

2.5 Logistics of Mastery Implementation at University of Illinois

In each of the following chapters discussing mastery-style exercises at the University of Illinois at Urbana-Champaign, the exercises were delivered through the online homework platform FlipItPhysics, which allowed levels[3] to remain locked until a threshold score on the previous level was achieved. Students worked until they submitted all answers, then the system graded the set of questions. After grading, multimedia correctives (in the form of conceptually grounded worked solution videos) were available for all questions, regardless of student correctness, and were available to be watched as many times as desired. Once a student was satisfied, they could call up a new version of the same unit (if mastery was not reached) or move on to the next unit (if mastery was shown). Calling up a new version of the same unit removed access to the previous version and solutions, so students had to attempt the new version without prior solutions at hand. However,

[3]From this point on, "units" of content described by Bloom and Keller are referred to as "levels."

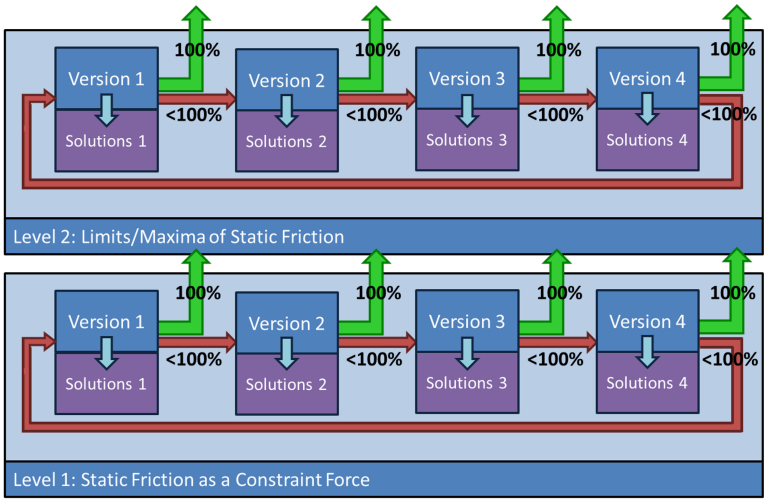


Figure 2.2: A schematic of mastery-style exercise format, including multiple versions of each level (which contains a number of individual questions), and each exercise typically consisting of a few levels, which open upon mastery of the previous level.

the last version seen of each completed unit and its solutions remained accessible after students moved on. Figure 2.2 shows a schematic for our local mastery implementation.

Chapter 3

Clinical Trial: Evaluating Correctives

3.1 Introduction

Though external research demonstrated that mastery can be an effective tool, it also warned that specific implementation affects its success. With the intention of eventually applying mastery-style homework to Physics 100, it seemed prudent to first get comfortable with the logistics of our specific implementation on a small scale, as well as check for an effect locally before investing more time in the project. Rather than implementing mastery for an entire semester, we ran clinical studies which focused on two topics that are notoriously difficult for students in the University of Illinois's introductory electromagnetism course: Gauss's Law and Electric Potential. The first of these studies was conducted in Spring of 2014.[1]

A major goal of the first clinical trial was to gauge the effectiveness of our correctives, which were narrated animated worked solution videos. The format was consistent with multimedia principles [66] employed in other learning materials at the University of Illinois at Urbana-Champaign. Multimedia learning tools with narration and animation have been shown to be equally or more effective than traditional study [5, 67]. Mayer's assertion is that people learn via auditory and visual channels; using visual input or audio input alone is less effective than employing both channels, but he also asserts that too much visual input can overwhelm the visual channel. Animations with little text and with audio input are preferable to animations with lots of text and audio input, and audio is preferable to audio with redundant text [68, 69]. The narrated animated worked solutions used minimum text, usually equations, and delivered the majority of explanation via the narration to best take advantage of multimedia principles.

The narrated animated solution videos further take advantage of the worked example effect [70]. The worked example effect is a consequence of cognitive load theory; working memory is taxed by problem solving, but removing some of the randomness of the process allows a person more cognitive power to focus on acquiring skills. The worked example effect is most useful for novices, as they lack the same schema expertise in the domain of the problem as more experienced learners. In fact, too much instruction for experts may add additional strain on working memory, a phenomenon known as the expertise reversal effect [71]; the

[1] A summary of this clinical trial was published as "Narrated animated solution videos in a mastery setting" in 2015 [64]. Much of the analysis was led by Noah Schroeder and is also included in his thesis, "Mastery-style exercises in physics" [65].

benefits of worked examples is diminishing with increasing expertise. Reisslein et al discovered that studying a worked example *before* working a problem is most effective for novices, whereas higher-level students gain more benefit from studying a worked example *after* problem solving [72]. Mastery-style learning allows students to watch narrated worked solutions when useful, before attempting the next version's problems, or after attempting the current version's problems, making the delivery flexible for needs of different students. The expertise reversal effect has been used to promote adaptive learning in general, advocating for learner-tailored instruction [73].

As far as feedback in general, Heckler and Mikula studied types of feedback in a similar context, using mastery-style exercises for vector math drills [62]. They discovered that knowledge of correctness paired with "elaborated feedback," such as an explanation or the correct answer provided the best benefit to both students with low and high prior knowledge. This was consistent with a meta-analysis of feedback in computer-based learning by van der Kleij et al, which found that the largest effect sizes came from elaborated feedback, compared to correctness information or knowledge of the correct answer [59]. The meta-analysis also found that students receiving immediate feedback (as opposed to delayed feedback) was more effective for lower level learning outcomes, and vice versa for higher level learning outcomes [59]. Computerized mastery homework favors immediate feedback because of its iterative nature, which should support building a strong foundation of understanding that is built early in students' exposure to content, as recommended for lower level learning outcomes.

3.2 Methods

The clinical trial was targeted to students in the calculus-based introductory electricity and magnetism course, typically taken by second year students. After the instructor advertised the study as a way to prepare for their upcoming exam, all students in the course were invited to participate via email and were randomly assigned to one of three groups: a mastery group, single try group, and control group. Students arrived in person for a three-hour session, which took place after they had completed the Gauss's Law lecture and homework and the electric potential lecture, but before completing the electric potential homework. Students were also promised an extra office hour following completion of the clinical trial to incentivize participation. The session was held on a Saturday, 10 days before the students' midterm exam.

The treatment for students in the mastery group saw two levels of mastery exercises. As recommended in mastery theory, small, specific competencies for each unit of instruction were defined. Our first level of instruction focused on Gauss's law in a planar geometry; students should be able to calculate the magnitude of an electric field around sheets or conducting slabs of charge, and conversely be able to find the charge density of a charged sheet, given the electric field. The second level's competencies targeted electric potential in a spherical geometry; students should be able to find electric potential differences over spherical shells of charge and conductors, or work from an electric potential difference between two points to find the charge density on a shell or sphere. These competencies were clearly defined and intentional. From these specific objectives, four extremely paralleled versions of sets of questions were created to test these competencies at each level, which were uploaded into students' online homework system. The mastery threshold to move from one level to the next was set at 85%. The first level (planar Gauss's Law) had a total of 8 questions, while the second level (spherical electric potential) had 9, so 85% translated to a leniency of one incorrect

answer in each level. Narrated animated worked solution videos were provided for all students, regardless of correctness on problems. The students were not forced to watch the videos, but were given access until they felt confident enough to try an isomorphic version of problems. Once the students initiated a new version, the process repeated until they had moved on to the electric potential level, where they repeated the process again, until they had mastered that level as well. The students in the single try group used the same materials as the mastery group, but were only given one version of each unit, whether they achieved the 85% threshold or not, and had access to the solution videos after completing their single sets, but were not required to watch them. Our control group was given no treatment at all.

Following their treatments (or non-treatment, in the case of control), students from all three groups took an identical written assessment, which asked them to find electric potential differences in planar geometries, as well as some conceptual questions, which included potential differences with spherical geometry. The assessment intended to synthesize the skills the students could have just learned: the ability to specify an electric field around planar geometry and to find electric potential difference by integrating over an electric field. The more transfer conceptual questions were intended to see how repetition of problem solving and solution viewing may affect broader conceptual understanding. The written assessment contained 10 problems, and four separate graders gave each problem a score between 0 and 3, meaning the assessment score ranged from 0 to 30. The grading was done without knowledge of which treatment group a student was in, and discrepancies larger than 1 point were discussed between graders. Once those discrepancies were resolved, each student's score for each question was the average of the four graders' scores. A schematic of the activities for the three groups is shown in Figure 3.1. The treatment problems and written posttest are all found in Appendix B.

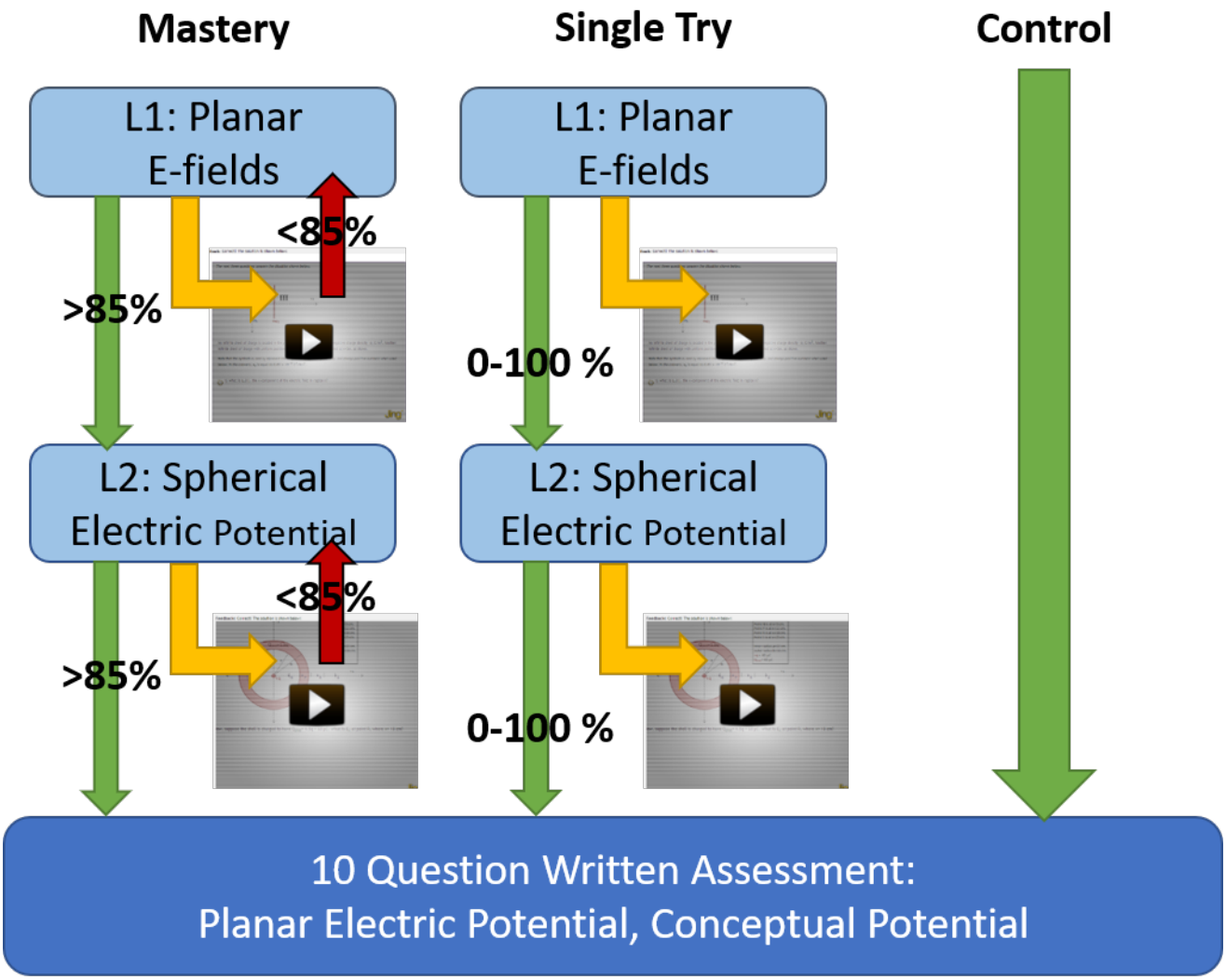


Figure 3.1: Treatments and assessment of three groups of students, where students were randomly assigned to a group. Students in the mastery group had access to videos (regardless of correctness) after submitting answers to a version, but were required to try repeated isomorphic versions until they had reached the 85% mastery threshold before moving on to the next level. Students in the single try group attempted the first version of each level and had access to videos (regardless of correctness) but were not required to make further attempts to access the next level, regardless of performance. Students in the control group had no treatment. All three groups completed a written assessment as a posttest.

3.3 Results

Of the original 88 participants who signed up for the clinical trial, 33 participants were placed into the mastery group, 33 were placed in the single try group, and 22 were assigned the control group. Within the mastery group, 24 participants actually attended the session, but only 17 finished the entirety of the materials. Of the seven who did not complete all the materials, 3 left without finishing the mastery exercises, 1 finished the mastery exercises but declined the posttest, and 3 left the posttest without finishing. Likewise, for the single-attempt group, 23 students actually attended the session; all completed the treatment exercises, but one did not finish the posttest. In the control group, 12 students attended the session and all finished the posttest.

3.3.1 Performance on Mastery Treatment

On the superposition set, all 24 students in the mastery group completed the exercises. On the students' first attempt (prior to receiving any of the learning material), the average score was 67%, and the average score on each of the 8 questions is shown in Figure 3.2. Students were familiar with this material, having already completed lecture, homework, and discussion (recitation) section before the clinical trial. The distribution of student scores on both the first and second attempt are shown in Figure 3.3. Nine students mastered on their first attempt (scoring 7 or 8 points of 8 to achieve the 85% mastery threshold), so the second attempt scores only include the 15 students who were required to redo the set. On their second attempt, 11 of the 15 students mastered the superposition level, and the last 4 students mastered on their third attempt. For students who did not master on their first attempt (15 of 24), improvement by question from the first to second attempt is also shown in Figure 3.4. Overall, these students improved from a score of 53% on their first attempt to 88% on their second, a gain of 35%, statistically significant with $p < 0.001$.

Similar analysis of students' performance and progression was done for mastery students in the following level testing electric potential. Figure 3.5 shows the average score on each question for mastery students on their first attempt of the electric potential set, where they averaged 44%. One of the students from the superposition set did not begin the electric potential level due to time constraints, so this plot represents 23 students' scores. There were three types of questions in the potential difference sets, which can be broken down to better understand students' incoming expertise with the material. Questions 1-3 asked students to calculate potential differences between hypothetical points, given functions for the electric field (e.g. $E(x) = 3x\,\hat{i}$); students scored an average of 17% on these three problems. The next three, Questions 4-6, asked students to find the electric field in a spherical geometry; the students scored an average of 83% on these problems. The last three problems, Questions 7-9, asked the students to find the potential difference around spherical geometries, combining the skills from the first and second types of questions, and students averaged 33%. Their higher score on Question 7 (48%) may have been due to equation hunting; kq/r successfully predicted the correct answer in Question 7 but did not calculate the correct answers for Questions 8-9, where students performed less well. Students' performance on these different skills makes sense in terms of prior exposure to material. As mentioned previously, the clinical trial took place after they had completed the unit for Gauss's Law and electric fields, but before they had much practice with potential difference. They did well finding electric fields but did not have expertise yet in using electric fields to find potential differences.

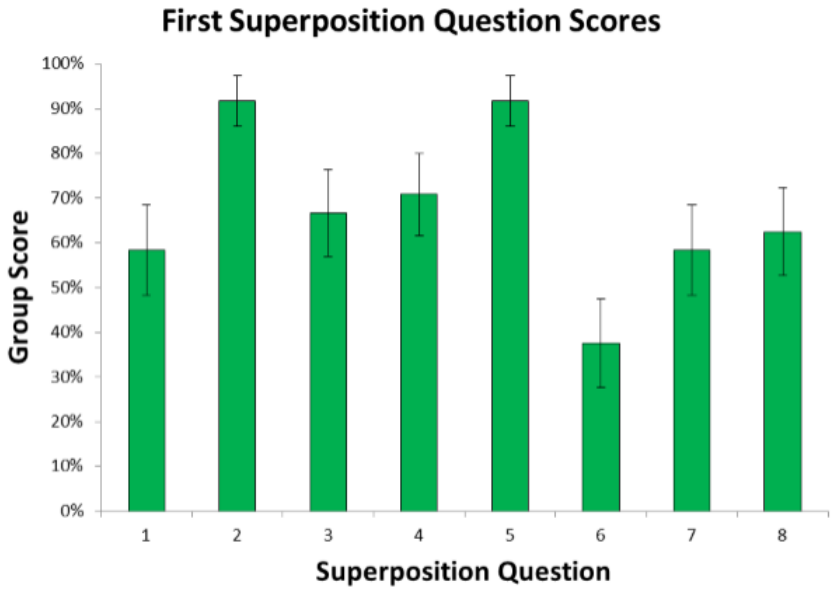


Figure 3.2: Average scores of 24 mastery group students by question on their attempt at the superposition level.

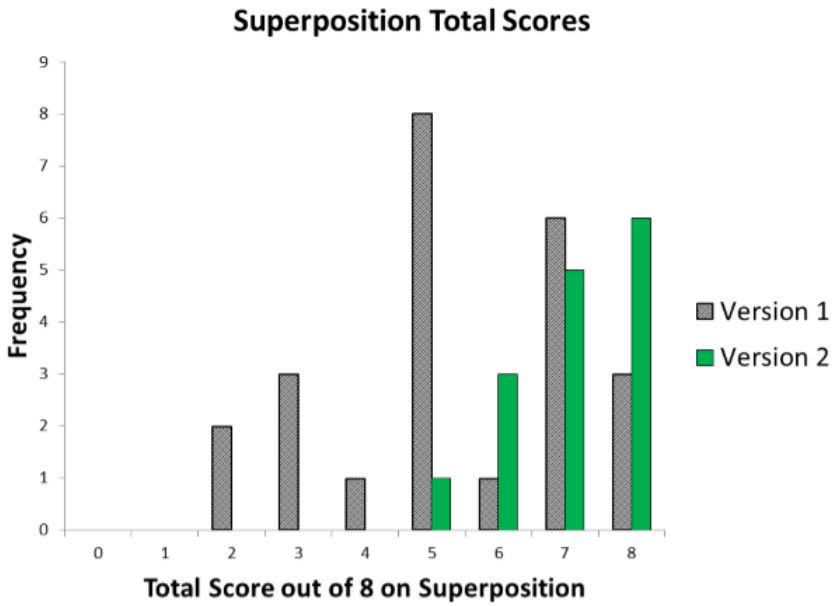


Figure 3.3: Histogram of mastery group student scores on their first and second attempt at the superposition level. The first attempt (Version 1) was done by all 24 mastery group students, but people who achieved mastery (a score of 7 or 8) on the first attempt did not make a second attempt, so are not included in the second attempt scores (Version 2).

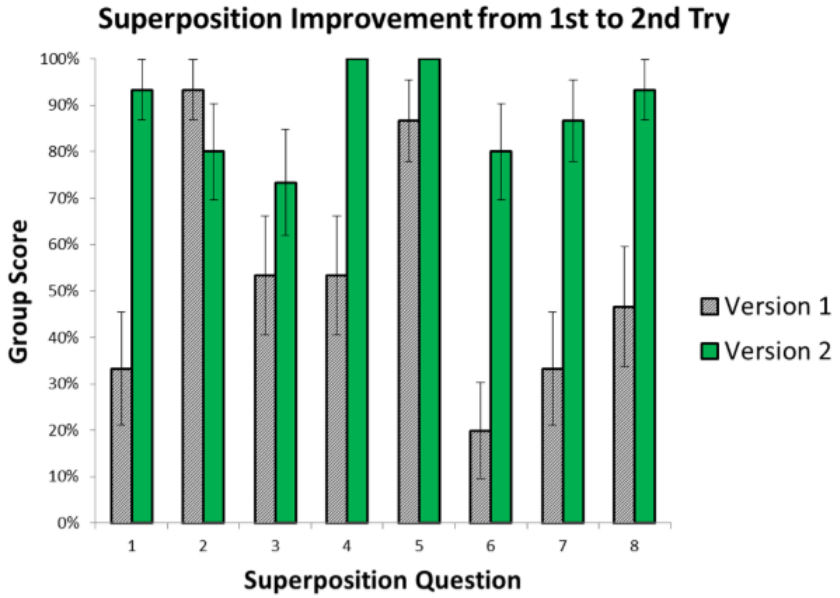


Figure 3.4: Average scores by question on mastery students' first and second attempts at the superposition level. These averages only include students who did not master the level on their first attempt.

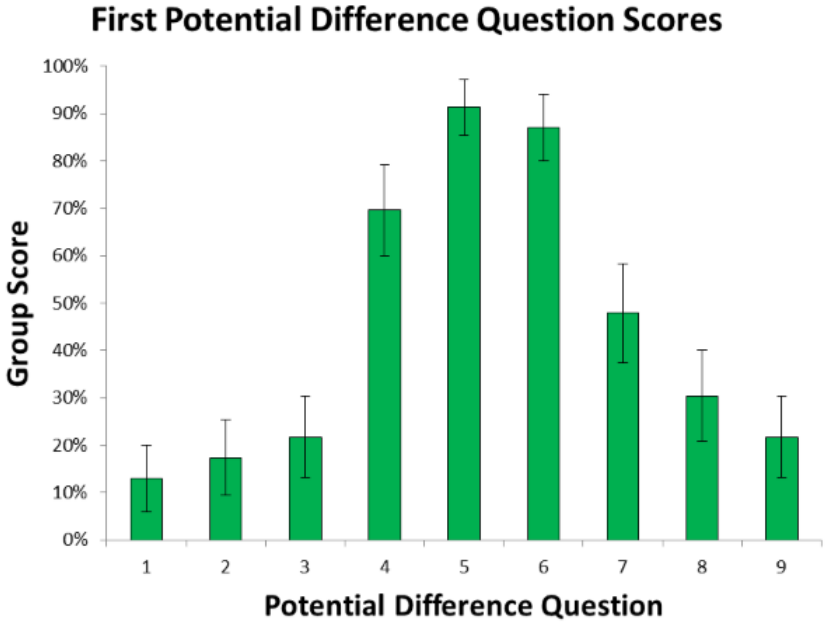


Figure 3.5: Average scores of 23 mastery students by question on their first attempt of the electric potential level.

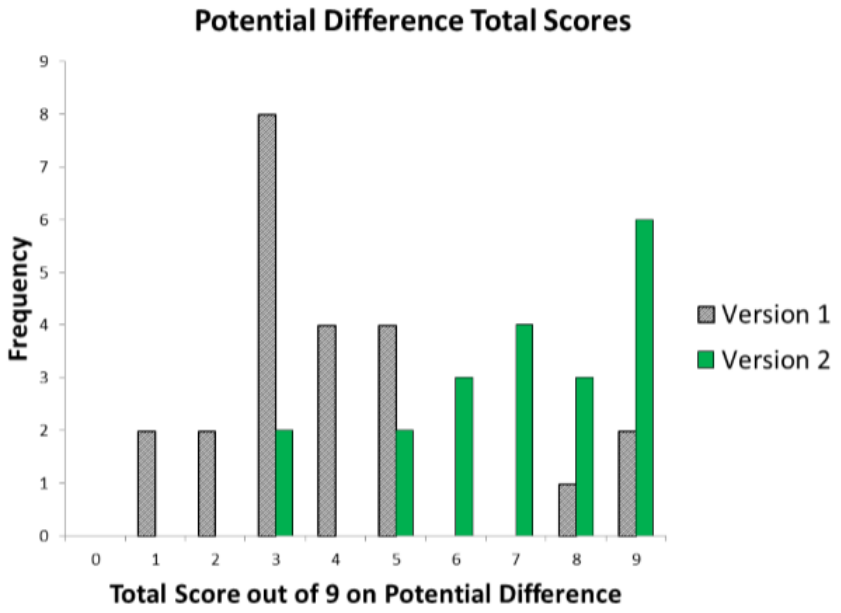


Figure 3.6: Histogram of mastery group student scores on their first and second attempt at the electric potential level. The first attempt (Version 1) was done by 23 mastery group students, but people who achieved mastery (a score of 8 or 9) on the first attempt did not make a second attempt, so are not included in the second attempt scores (Version 2).

Table 3.1: Summary of average scores by mastery group students on attempts 1 and 2 (versions 1 and 2, respectively) for students who did not master on their first attempt (and thus, had an opportunity to demonstrate improvement through a second attempt).

			V2 improvement from V1	
	Version 1	Version 2	Gain	p-value
Superposition	$53 \pm 4\%$	$88 \pm 3\%$	35%	<0.001
Potential Difference	$37 \pm 3\%$	$78 \pm 5\%$	41%	<0.001

Likewise, Figure 3.6 shows the distribution of student scores on the potential difference problems from their first to second attempt. Only three of the 23 students mastered on their first attempt (a score of 8 or 9 out of 9 questions), so there are 20 students represented for the second try (Version 2). Considering the 20 students who required at least two attempts, there was statistically significant improvement from the first to second attempt, an average score of 37% to 78%, a gain of 41% with corresponding p-value of $p < 0.001$. Each question showed improvement, shown in Figure 3.7, with 7 of 9 being statistically significant, and the other two questions' improvements saw a ceiling effect. Nine students mastered the electric potential content on their second attempt, seven mastered on their third attempt, one on their fourth, and three had to leave before they finished the material. Students seeing the mastery versions saw significant improvement on both levels, and their scores are summarized in Table 3.1.

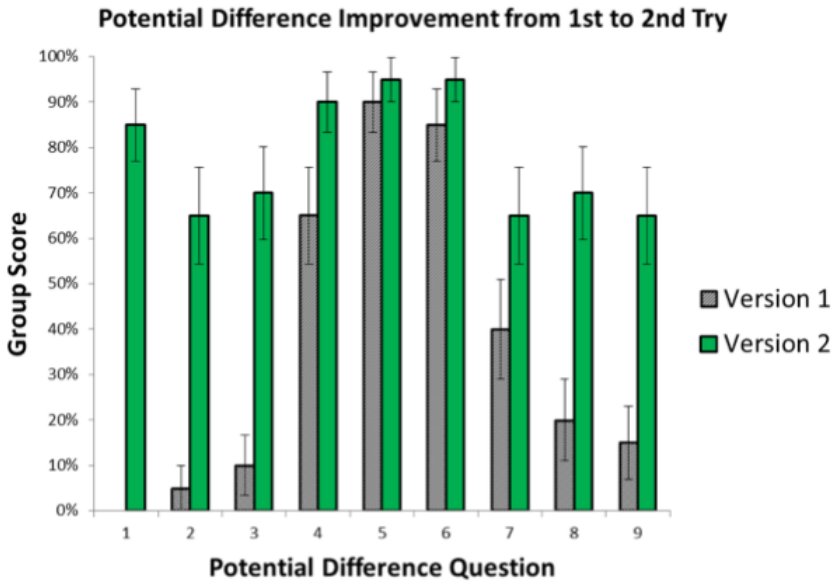


Figure 3.7: Average scores by question on mastery students' first and second attempts at the superposition level. These averages only include students who did not master the level on their first attempt.

3.3.2 Student interaction with solution videos

Between attempts of the different versions of mastery, students' use of the solution videos could be measured using time stamps in the online system. By taking the difference between the time of their submission to grade the current set of questions and the time to call up the next attempt, an estimate of time they spent watching videos was calculated for each student. This technique also allowed us to know how long the single try students spent watching videos between the superposition level and electric potential level. There is some uncertainty in the exact time, as students spent some time reviewing their answers between tries as well, but the majority of their time was spent reviewing videos and taking notes on the videos. Anecdotally, students were diligent in watching and worked efficiently through the material. There were few incentives for unproductive time because students were participating voluntarily and were able to leave (or receive extra time with tutors) when they had finished. Figure 3.8 shows the scatter plot of students' time spent watching solutions to superposition level problems relative to the number of questions they missed on their first attempt.

Unsurprisingly, students who missed more questions spent more time watching videos. Particularly, those who missed only one question spent only a few minutes watching, which is around the length of each video. Both the mastery and single group students spent considerable time watching videos, which was unexpected; we anticipated that the students with the single try condition would not have incentive to view solutions without the pressure of needing to master a successive try. The behavior of the single try students suggests that students found value in the videos, with or without external pressure for mastery. From observation during the session, most students worked eagerly with the narrated worked problems, regardless of their group. For the electric potential set, a similar scatter plot is shown in Figure 3.9. Because our measurement of solution time was based on the time stamps for completing one version and calling up the next, those who

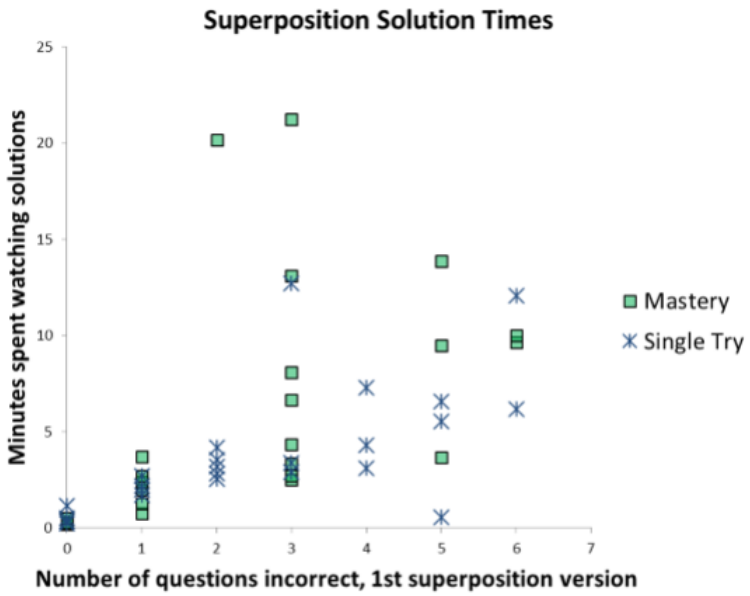


Figure 3.8: Scatterplot of minutes spent with superposition level solution videos by students in the mastery and single try groups, based on the number of questions missed on their first attempt at the level. For mastery students, r = 0.25 and for single try students, r = 0.64.

had no online activity after the first attempt at the electric potential level are excluded from this plot; we do not have solution time estimates for students who mastered on their first try or for the single try students. Within the mastery group, however, there is also a positive correlation between time spent with solutions and the number of questions missed on students' first attempts and a considerable amount of time being spent.

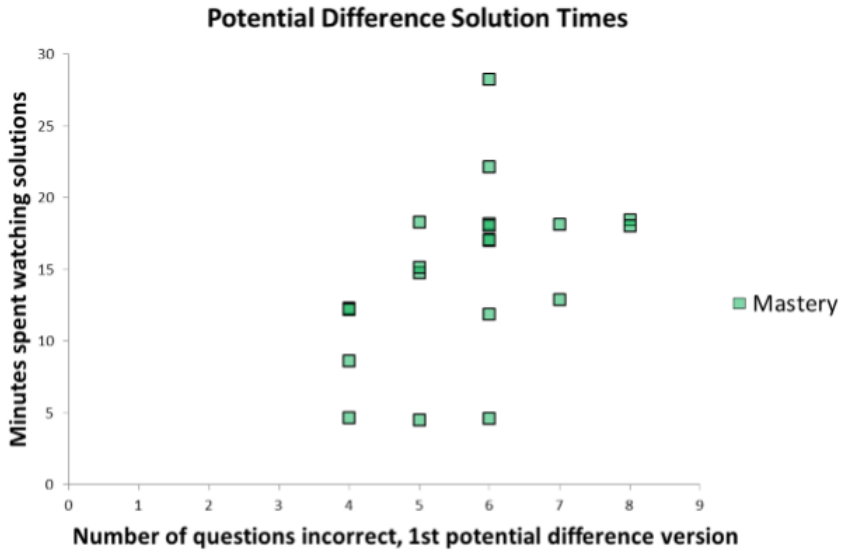


Figure 3.9: Time spent with electric potential level solution videos by students in the mastery group, based on the number of questions missed on their first attempt at the level. This plot does not include students who mastered on their first try or students in the single try group, as they did not have another version to call after completing their first set, and thus, no ending time stamp to calculate estimated time with solutions.

3.3.3 Posttest Performance

Following their different treatments (and no treatment by the control group), the three groups completed a written assessment. Recall that not all students finished the treatments in the time they had allotted themselves for the activities and left before finishing the posttest. Of the original 24, only 17 students in the mastery group completed all the materials. Likewise, 23 of 24 students in the single try treatment finished the activities and the posttest. All of the students in the control group (12 students) finished the posttest. The results discussed are the subset of the original students who were able to complete all of the activities, which are shown in Table 3.2.

For students in the control group, the posttest was very difficult, with students averaging about 4 of 30 points. Because the students in the control group received none of the treatments, their scores can be considered a proxy for a pretest to the treatment conditions, which both scored significantly higher on the posttest than the control students. The difference between the mastery students and single try students

Table 3.2: Posttest Scores (out of 30) for each treatment group, with statistical comparisons of performance across groups.

		Compared to Control		Compared to Single Try	
	Posttest Score	effect size	p-value	effect size	p-value
Mastery	18.4 ± 1.4	2.9	<0.0001	0.24	>0.2
Single Try	16.9 ± 1.4	2.4	<0.0001		
Control	4.1 ± 1.1				

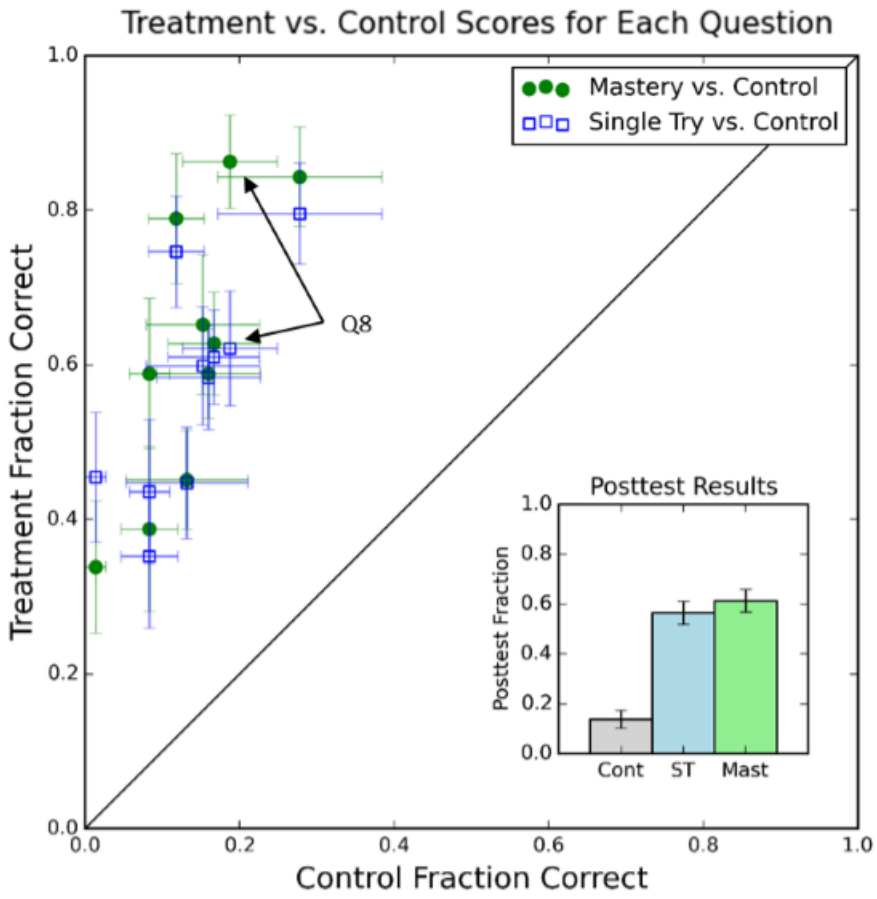


Figure 3.10: Outset: Scatterplot by question on posttest, showing performance of treatment groups compared to control. Points above the diagonal line show higher performance by students in the treatment. Question 8 is indicated, the only question that was near transfer to the treatment materials. Inset: Overall score on posttest by group.

was not statistically significant, likely because the mastery and single try students behaved so similarly while interacting with the solution videos. A question by question scatter plot shows gains on the posttest compared to control, in Figure 3.10. Students in the treatment groups scored significantly higher than control for every question. Only one question on the posttest was near transfer to the materials in the treatment exercises (Question 8, indicated on the plot), and this question had the largest difference between the mastery and single try groups, with $86 \pm 6\%$ and $62 \pm 8\%$ averages respectively, giving an effect size of 0.75[2] and $p < 0.05$. This problem was likely easier for mastery students who had performed more repetitions of a similar problem.

[2]Calculated effect sizes in this thesis refer to Cohen's d effect sizes.

3.4 Discussion

On the mastery exercises, students showed consistent and significant improvement from one attempt to the next on both subjects, one which was familiar and one which was early in their exposure to the content. This indicates that the act of repeated testing coupled with the narrated animated worked examples was effective for improving their learning and giving them an opportunity to demonstrate that learning, both on new and practiced skills. Compared to students in the control group, those in the mastery and single try treatment both outperformed on the written assessment. We expected to see a difference between the mastery and single-attempt groups, but their performances were remarkably similar, possibly due to their similar behaviors. Although we expected only students with the pressure of the mastery requirement to watch solutions, the single-attempt students worked diligently and consumed all of the content available to them, most likely in light of their upcoming exam. The majority of mastery students who did not master on their first attempt ended up mastering on their second, suggesting that one set of solution videos (which both groups saw) was sufficient for many students and explains the similarity between the groups' performance. Thus, the real comparison was between the treatment groups (mastery and single try) and control group, which showed the learning effects from the solution videos and overshadowed the effect of the delivery method.

The solutions were well received by students, demonstrated by their diligence in watching, particularly for the single try group. Spending more time with solutions meant that students had less time to spend with tutors after the activities were finished, but the group who had no stake in getting another set correct still chose to watch. Students who missed more problems generally spent more time with solutions, suggesting that they found the videos helpful for correcting misunderstandings or learning new skills. Though there was pressure to do well on the following set for the mastery students, watching the videos was completely optional for everyone.

On the written posttest, only one question was near transfer to the treatment activities. Most were a synthesis of taught skills, but not direct isomorphic problems to examples they had seen. This is encouraging, particularly in light of the criticisms of mastery for teaching algorithmic skills only [46]. Students who saw the treatments were able to synthesize and apply what they had learned to new situations and conceptual questions that were not addressed in the treatment.

Although students responded well to the videos and activities, observation during the session let us see students' fatigue and frustration. Some complained, and some were not able to finish the activities in the three hours allotted. We allowed students to stay as long as they needed, but many had other commitments. This session was time they had set aside, but students working at their own pace at home may not have the same determination without the promise of a reward (tutors), and would likely not work through all the problems in one single session. There is literature supporting distributed practice over a single study session [74–76], but many students tend to complete their homework in one sitting, and we would like the materials to be achievable. Sets of problems that are 8 or 9 questions long may make the stakes for error too punishing, especially because there are multiple sub-skills in each level. We attempted to reduce the stakes for small errors (such as rounding) by making the questions multiple choice, but keeping track of many small steps in a problem is demanding on cognitive load. Even with high skill, requiring chains of small details in each problem across many problems creates a high probability for one error that can require students to redo an entire set of 8 or 9 problems.

Goals for a repeated experiment would be to create more distinction between the mastery and single try groups. The main takeaway from this study was the effectiveness of the videos in helping students learn, but we could not distinguish whether the mastery system as a whole was more effective than other delivery systems. In both our treatment groups, students were able to take advantage of many of the learning principles which make mastery work; all treatment students received frequent elaborated feedback [61], the feedback was given in time with students' work [31], and all treatment students had access to solutions that took advantage of multimedia principles and the worked example effect [66, 70]. Moving forward, we could be confident that our solutions were effective correctives and focus next on the delivery system compared to existing online homework.

Chapter 4

Clinical Trial: Evaluating Mastery Delivery

4.1 Introduction

4.1.1 Online Homework Background

In the last decades, online homework in physics courses has become standard for many universities. A number of commercial online homework platforms exist for widespread use across institutions. WebAssign, one of the largest platforms, can be used for a variety of courses (including physics) and is used at over 2,600 institutions, serving over a million students [77]. Pearson also has a online homework series called "Mastering" which includes "Mastering Physics" [60] that is widely used. Despite its name, Mastering Physics does not use mastery learning as defined in this dissertation. Promotional materials for WebAssign and Mastering Physics include data-driven case studies showing effectiveness of the online homework systems [78, 79]. Though many are not peer reviewed, one case study for WebAssign corresponds to a doctoral dissertation [80]. There is more literature on the effectiveness of smaller systems designed by researchers, such as LON-CAPA and OWL [54, 81–86].

The general effectiveness of online homework, particularly in contrast to paper and pencil homework, has been evaluated over the last decades [86–89], including literature reviews of online homework [90, 91]. The most conservative finding is that online homework does not hurt students' learning, but provides many administrative benefits by offloading grading for large classes where homework where may not otherwise be feasible [88, 92], while others show gains in performance for a variety of types of students using web-based homework [85, 86, 93]. Those that find equivalence between online systems and paper and pencil homework cite the importance of grounding homework in learning principles; being online itself is not inherently better, but one must be intentional in its use to take advantage of its properties. Thoennessen et al cite the advantage of online homework for faculty to understand students' behavior and difficulty with different types of problems using statistics [85], while Kortemeyer favors online homework for its capability to incorporate multimedia learning [91]. For example, the addition of multimedia learning modules has been shown to improve students' learning and exam scores, and can be implemented alongside online homework [1–6]. Bugbee and Bernt also showed that students spent significantly more time completing online testing than paper and pencil testing, while the students believe that they spent less time [94], meaning that students

were happier with the online system while also getting the advantage of more time on task. An analysis of time spent on homework by Tang and Titus also showed that students tended to spend more time on their homework using an online system than when doing homework on paper [95]. Generally, the amount of time on task spent by a student correlates with their learning [96], and online homework's interactive and responsive nature can engage students for more time with less burnout or boredom [95].

Another major feature of online homework is its ability to give responsive feedback in real time. Besides eliminating the restraints of time and personnel that come with hand grading paper assignments (particularly for large courses), online homework has the ability to give students personalized feedback while they are working. There is conflicting research about the effectiveness of immediate versus delayed feedback [97, 98], while a review of formative feedback by Schute suggests that different types of skills are better suited to each timing of feedback [99], namely that higher transfer tasks should have delayed feedback while more simple or procedural tasks favor immediate feedback. For a series of questions, Butler and Roediger found that students had better recall from getting feedback on exams after completing the whole exam, rather than receiving feedback for individual questions as they worked through it [100]. Most online homework systems allow different settings for the timing of feedback, rather than feedback being constrained to the schedule of the instructor. Types of feedback can also vary; Heckler and Mikula found that pairing correct/incorrect feedback along with an explanation was superior to only correct/incorrect feedback [62]. In early studies of web-based homework, immediate feedback was criticized for being less thorough than written feedback [89], but current technology now lets instructors give students more information than correctness, often tailored to students' individual mistakes and grounded in conceptual explanations. Online homework systems are flexible and allow instructors to cater their feedback timing and content to the topic and their personal preferences. As stated before, online homework is not inherently better, but can be a powerful tool when instructors use literature to guide its use to take advantage of learning principles.

4.1.2 Online Homework at the University of Illinois

In the physics department at the University of Illinois at Urbana-Champaign, the first computer-based physics lessons were through a system called PLATO, implemented in the mid 1970s [101, 102]. These computer-based activities were used for a third to a half of students for many years and required terminals for use. Starting from 10 experimental terminals, a grant from the National Science Foundation allowed the department to expand to about 30 permanent terminals, shown in Figure 4.1. By 1975, over 500 students per semester were using the PLATO classrooms. A developer of PLATO, Dennis Kane, was also involved the next iteration of computer-based online homework, CyberProf, which was developed by Alfred Hübler [103, 104]. Cyberprof began being used in the physics department in 1995; by 1997, 16 courses were using the system, at the University of Illinois and other institutions [103]. Cyberprof then led to an online platform called Tycho [105] in the late 1990s, which allowed the integration of interactive examples. The interactive examples are problems developed at the University of Illinois which have optional scaffolding for students, beginning from conceptual analysis, then moving on to strategic and quantitative analysis. The use of scaffolding reduces cognitive load for students who request help but avoids the expertise reversal effect [106] by being opt-in; those who know how to do the problem are not required to interact with the scaffolding.

Tycho evolved into smartPhysics, now rebranded as FlipItPhysics, and is currently used by students in physics courses at the University of Illinois. The majority of homework in introductory physics courses at Illinois

Figure 4.1: An early classroom with 32 PLATO terminals [101].

uses a combination of immediate and delayed feedback questions via the online system, still incorporating interactive examples for some difficult multi-step problems. The feedback provided to students is answer correctness, and in cases of common conceptual errors, wrong answer feedback is programmed in that can point out the conceptual errors that lead to the incorrect answer that was chosen. The mastery delivery system is also run through FlipItPhysics, and gives immediate feedback as a set after the entire set of problems is completed, as suggested by Butler and Roediger [100] and mastery literature in general. As mentioned in the previous two chapters, the system prevents students from accessing the next level of problems until they master their current level via repetitions of multiple versions of the level. FlipItPhysics allows the worked solution videos to be embedded into the system so students can access the solutions in the same place as they work their problems. As a response to the questions left unanswered in the Spring 2014 Clinical Trial [64], another clinical experiment was done in Fall 2014 to evaluate the mastery delivery system itself, which is documented in this chapter.[1] There are many methods of delivering online homework to students, and this clinical trial's purpose was to evaluate the mastery method's effectiveness compared to the traditional online homework at the University of Illinois.

4.2 Methods

Like the previous clinical trial, this experiment was advertised to students in the introductory calculus-based electricity and magnetism course for engineers, the second course in the engineering physics sequence which is typically taken by students in their second year. The study was again advertised by the instructor via email, with the incentive of access to experienced tutors following the clinical trial. The timing of the

[1] A summary of this clinical trial was published as "Clinical study of student learning using mastery style versus immediate feedback online activities" in 2015 [107]. Much of the analysis was led by Noah Schroeder and is also included in his thesis, "Mastery-style exercises in physics" [65].

clinical trial was after students had completed the unit on Gauss's Law and electric fields and had attended lecture for the electric potential unit, but had not yet done the homework or discussion (recitation) sections corresponding to electric potential. As these two topics are notoriously difficult for students in our course, the study was again advertised as a way to prepare for the upcoming exam. The clinical trial had sessions on both Saturday and Sunday, which corresponded to 11 and 10 days prior to the first hour exam, which included both electric fields and electric potential. Two sessions were offered to give students more flexibility in scheduling, but each student attended only one of the two sessions. The same topics and questions were covered for the treatments (Gauss's Law for planar geometry, with an emphasis on superposition, and electric potential in spherical geometries) and the written posttest (electric potential in planar geometry and conceptual questions).

Some alterations to the delivery from our first study aimed to address its limitations, both the indistiguishability of the mastery and single try groups and student frustration. While the first clinical trial gave us evidence that the narrated animated solution videos helped students learn, we had compared the mastery-style delivery to a hybrid of mastery and traditional homework; both the single try and mastery groups behaved similarly and took advantage of the same learning principles. Our next objective was to discover if mastery-style homework was more effective than traditional online homework, which was not represented in the first clinical trial. Thus, instead of dividing students into mastery, single try, and control groups, we split students between mastery and traditional homework groups. Both treatment groups in the original study outperformed the students in the control group by a large, statistically significant amount, so the control was not needed for its second iteration, which improved our statistical power.

For students in the mastery group, there were some changes to the delivery of the content from the previous clinical trial. The previous threshold of 85% signaling mastery was raised to 100% to emphasize the importance of mastering each topic before synthesis, but levels were broken into smaller sets. The superposition material (which was originally eight questions as Level 1) was split into two smaller sets, focusing the competencies for each level. Questions 1-4 became Level 1, all of which asked students to find electric fields from given charge densities of sheets and slabs in planar geometry. Questions 5-8 became Level 2, which asked students to find charge densities given the electric field at some defined location. This was done in response to the higher mastery threshold demanded and student frustration from the first experiment. We hoped that splitting up the content may allow students to focus their practice and reduce the penalty of mistakes; redoing a level is less punishing if the levels are more reasonably sized sets. The nine questions from the electric potential level in the previous clinical trial were also split into two smaller levels: five questions that asked students to set up and solve a single integral became Level 3, and four questions which asked students to do multiple integrations became Level 4. This is a clear example where students must be able to do the current level consistently to ensure success in the following level; students who cannot set up and solve a single integral to find electric potential will struggle to solve multiple integrals over different regions. All of the questions were given in multiple choice format to battle false negatives due to rounding or mistyping. The wrong answer choices included conceptual distractors, which we considered true negatives, but we did not want students to be penalized by doing a set of problems over if they were competent in the skill already and had made a careless error.

The non-mastery group (single try) was re-defined to more accurately represent typical homework delivery in physics courses at the University of Illinois at Urbana-Champaign. The homework group saw the same problems as the mastery students, but were given immediate correctness feedback on each attempt, question

by question, rather than being forced to complete a whole set before receiving feedback. They could try individual questions over as many times as needed without penalty on their version of questions; only one version of each level was given to students in the homework group. The restriction of finishing the current level before attempting the second was also removed. Students saw no solutions to the problems (as is typical in their standard online homework), but were allowed any resources that would typically be used, such as notes, classmates, and the graduate students running the experiment (to simulate a TA). The homework group saw a combination of free response and multiple choice questions; all numeric calculations were asked for as free response, and questions with a conditional or symbolic answer were still logistically required to be multiple choice.

Following the treatments, the two groups took a common posttest, the same 10 question written assessment from the previous experiment. The posttest included questions about electric potential differences in planar geometry, to synthesize their ability to find electric fields around planar geometries with their ability to find electric potential differences, given an electric field. The questions also included conceptual tests of both planar and spherical geometries which were not addressed directly in the treatments. Two graders (both of whom graded the assessment in the previous trial) scored the assessment out of 30, with 0-3 points given for each question, without knowing which group each paper's owner belonged to. Question scores which differed by more than one point between the graders were discussed and resolved to be within one point of each other, and both graders' scores were averaged for analysis. A schematic of the treatment for the two groups is shown in Figure 4.2, and the materials and posttest are attached in Appendix B.

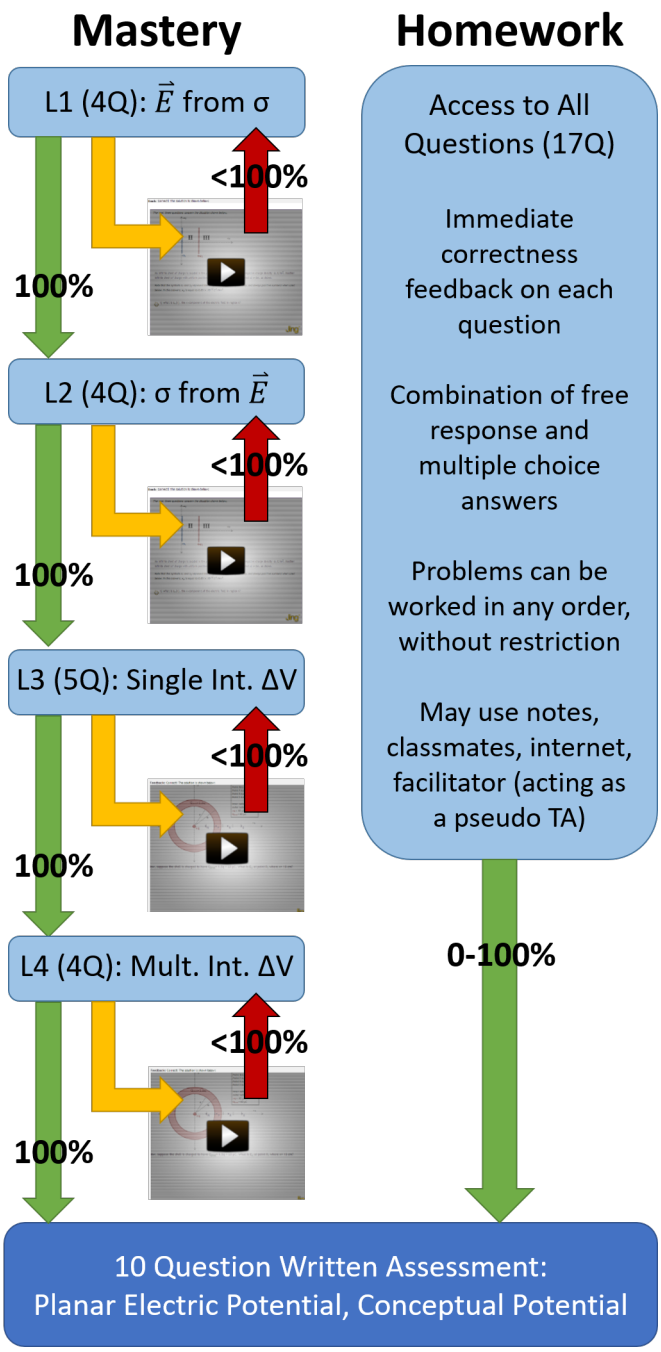


Figure 4.2: Treatments and assessment of two groups of students, where students were randomly assigned. Students in the mastery group had access to videos (regardless of correctness) after submitting answers to a version, but were required to try repeated isomorphic versions until they had reached 100% mastery threshold before moving on to the next level. Levels 1 and 2 corresponded to Level 1 in the previous clinical trial, and Levels 3 and 4 corresponded to Level 2, shown in Figure 3.1. Students in the homework group had access to all problems and received immediate feedback of the correctness of their answers for each submission to single questions, with unlimited attempts. They were also able to use reference materials, including other students, as they would typically do while completing online homework. Both groups completed a written assessment as a posttest.

4.3 Results

In evaluating the effectiveness of the different types of homework delivery, students' time spent, performance and progression through problems and levels, and posttest performance were evaluated. From the email invitation, 126 students signed up to do the experiment and were randomly split between the two treatment groups. Of those students, 32 from the Mastery group and 27 from the Homework group actually attended the experiment.

4.3.1 Time Spent on Treatment Materials

A potential two-sided disadvantage/advantage often cited for mastery is its extra demand on students' time compared to more traditional homework, which can be frustrating for students but has the benefit of increasing the learners' time on task [10, 96]. In this clinical experiment, we were surprised to see that students in both the Homework and Mastery groups had similar distributions of time spent, which are plotted by level in Figure 4.3 as a box and whisker plot. The plot shows the interquartile range (IQR) within the shaded boxes (the middle two quartiles), and identifies the median of each distribution as the dark horizontal black lines. The whiskers represent the extreme values, either the highest/lowest points in each distribution, or held at 1.5 $\times$ the IQR beyond the boundary of the IQR, if the maxima are larger or smaller than that threshhold. The distributions were skewed and did contain high outliers (larger than the the 1.5 IQR threshhold), so this type of plot is well suited to display the data and the median is more representative of student behavior than the mean for each group.

The median values of time spent on each level are nearly identical for students in the homework and mastery groups. This was surprising, considering that students in the mastery group were held to a higher standard of correctness and typically solved more problems, as well as used time to view solutions. It appears that the after-level feedback and worked example videos may have compensated for those additional requirements. Although both groups spent similar amounts of time, looking at student behaviors within that time reveals differences in how the students spent it.

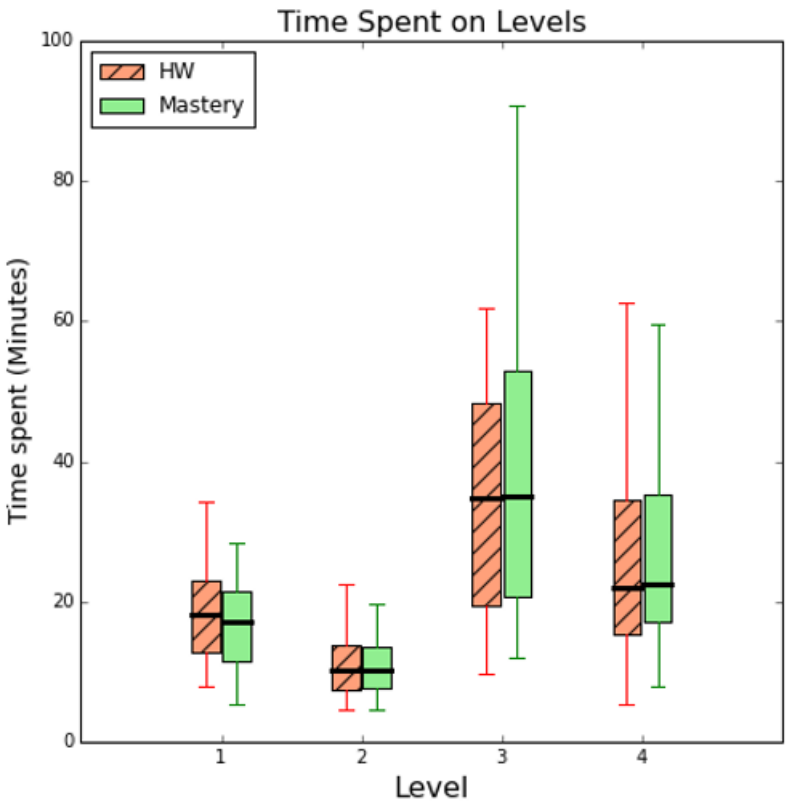


Figure 4.3: Box and whisker plot of the time spent by students in each treatment on each level. The black, horizontal lines represent the median time for each group, and the shaded boxes represent the interquartile range (IQR=Q3-Q1). The whiskers represent either the highest/lowest values of time spent for that group, or 1.5 $\times$ the IQR beyond the edge of the IQR, in the case where there were outliers beyond that value. The outliers are not shown on the plot.

4.3.2 Progression and Performance on Treatment Materials

Despite similar amounts of time on the material, students' methods of going through the material were different, which can be seen in the box and whisker plot of the number of submissions students averaged for each question within the levels. This plot is shown in Figure 4.4. This method of representation was again chosen due to the skewedness of the data and the number of upper outliers. As before, the median is represented by the horizontal bar, which shows differences between the two groups. The Mastery group had a lower median number of submissions for every level, with the difference significant on levels 2 and 4 ($p < 0.01$ via the Mann Whitney test, which was chosen because the distributions were not normal). This is unsurprising, since students in the homework group were not penalized (by getting a new set of questions) for wrong answers, and many resorted to trial and error. In an extreme example, one student had 93 attempts at a single question. When given feedback that their answer was off by a power of ten, students tended to try other iterations of their answer incrementally rather than going through their work and re-calculating. Trial and error was not a reasonable strategy for students in the Mastery group, as missing a question on the first submission meant that they were given a new set of questions, where their previous work was not relevant to finding the exact answer to the isomorphic new set. Some students in the Mastery group did "skip" the first version, submitting answers without spending time solving to go straight to solutions, but this is indicative of their ability to recognize themselves as novices and they were able to apply what they learned from the worked examples to different questions on the next set.

The differences in format delivery also informed some differences in student progression through the set or sets of problems. Figure 4.5 shows the average fraction of students correctly answering each question on their first attempt. Because students in the homework group were able to see feedback on questions before their first attempt on consecutive problems, they showed greater improvement question to question. For example, Questions 1 and 2 in Level 1 were similar; students in the Homework group did much better after they had received feedback and been able to work through Question 1 to correctness before attempting Question 2. The students in the Mastery group did not receive feedback until they had answered all questions in the set, and thus performed very similarly on the two similar questions. The same pattern is seen in Questions 1 and 2 in Level 4; Mastery students scored similarly on the two, but the Homework students showed improvement from the first question to the second.

In some questions (indicated by asterisks in Figure 4.5), the mastery students saw answer choices as multiple choice while the homework students entered their answers as free responses. It is possible that this was an advantage for mastery students; if they first calculated an answer that was not listed, they could try again. Investigating the content of the free response answers to attend to this difference gave increases in scores on four questions for the homework students (if we throw away any answers that are not listed as multiple choice options), but none that affected the general trends discussed. The largest gain was in question 2 of level 2, which increased to 58%, and other increases, which were much smaller, were in question 2 of level 3, and questions 1 and 2 of level 4.

For questions without format differences, such as questions 1, 3, and 4 in level 2, mastery students still significantly outperformed the homework students. Level 2 used many of the skills taught in Level 1, so their first attempt advantage suggests that the mastery students learned more from their work on the previous level. In level 4, questions 1 and 2 were very similar to the first question in Level 3, but homework students did not show improvement from the level 3 question into level 4, whereas the mastery students performed

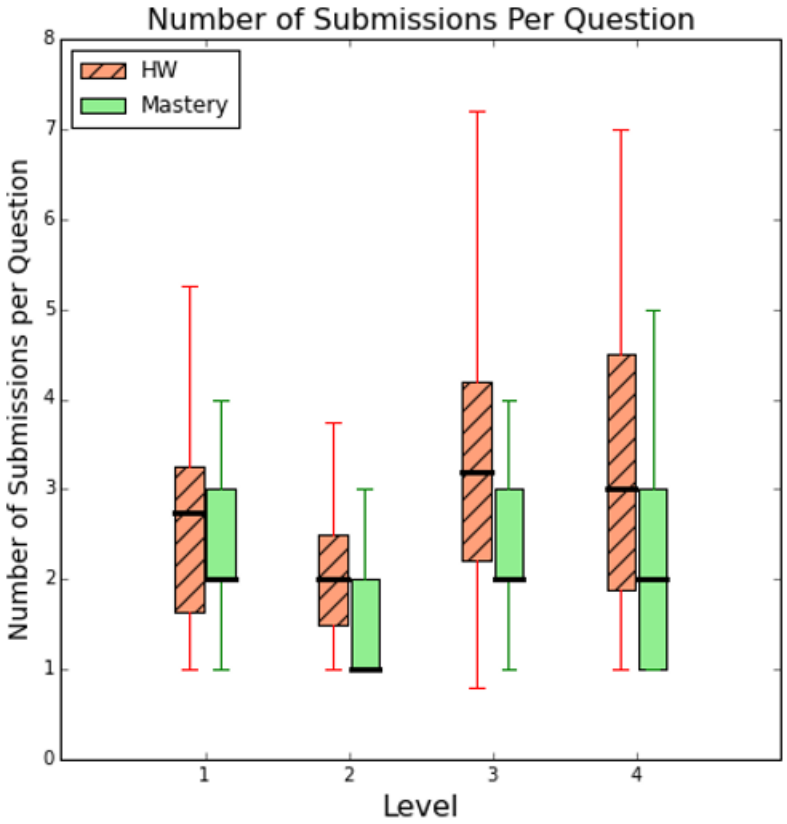


Figure 4.4: Box and whisker plot of students' average number of submissions per question within each level, based on their group. Horizontal black lines represent median values, which are within the shaded box representing the interquartile range (IQR=Q3-Q1). The whiskers represent either the highest/lowest values of time spent for that group, or 1.5 $\times$ the IQR beyond the edge of the IQR, in the case where there were outliers beyond that value. The outliers are not shown on the plot.

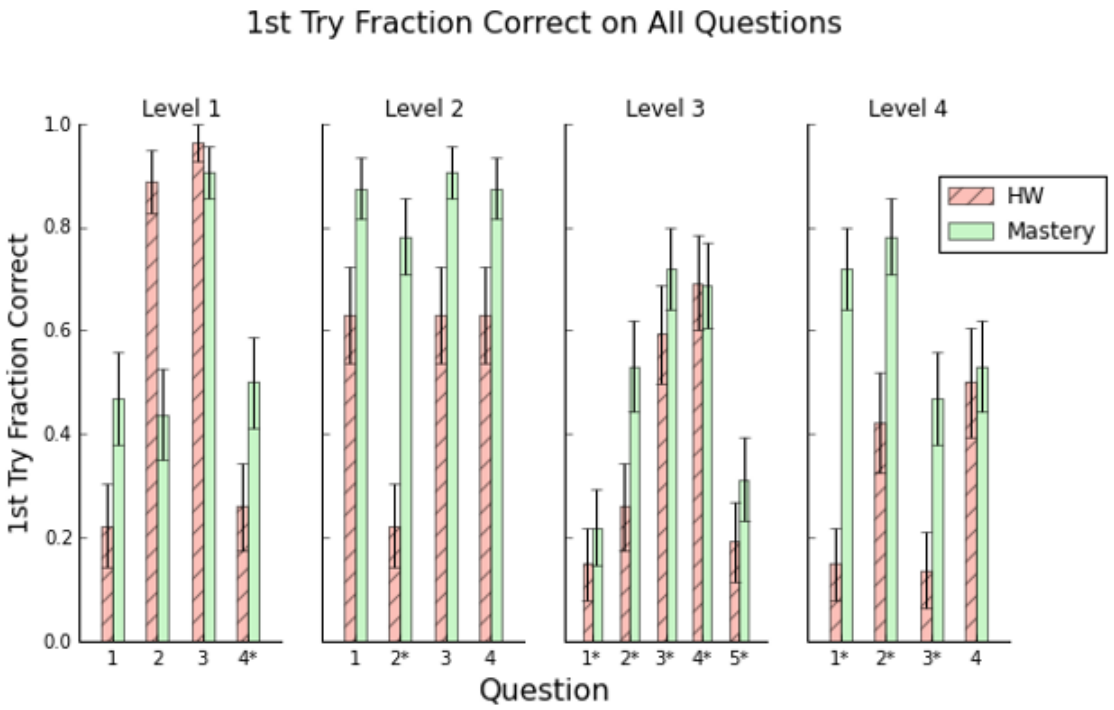


Figure 4.5: The fraction of students with the correct response on their first try on each question. Asterisks indicate that the questions had different modes of inputting answers; Mastery students always had multiple choice answers, but questions with asterisks were free response for Homework group students.

very well. Again, this suggests that the mastery students learned more from their work on Level 3, which they were able to apply to questions in Level 4, whereas the success by the homework students in Level 3 did not carry forward.

4.3.3 Posttest Performance

On the common written assessment, the improved performance of the mastery students persisted. Out of 30 points, the students in the mastery group scored an average of 16.0 ± 1.2 compared to the homework group's average of 9.7 ± 1.3, which was statistically significant with $p < 0.005$.[2] A plot of the overall score and scores by question is given in Figure 4.6. The transfer between the treatment materials and the posttest varied from near to far transfer by question. The most similar question was Question 8 on the posttest, which was very similar to question 3 of level 4 in the treatment materials, and showed the largest gain for mastery students over the homework group. The question required students to integrate over two regions, one inside a conductor, and relied on skills that were trained throughout Levels 3 and 4. The mastery students also outperformed the homework group when combining expertise of electric fields in planar geometry and electric potential in spherical geometries to find electric potential differences in a planar geometry (questions 2, 4, and 6). Question 5, where the two groups performed the most similarly, asked students to work backwards

[2]The difference between the scores was significant whether using either grader's scores or the average of the two, which is reported.

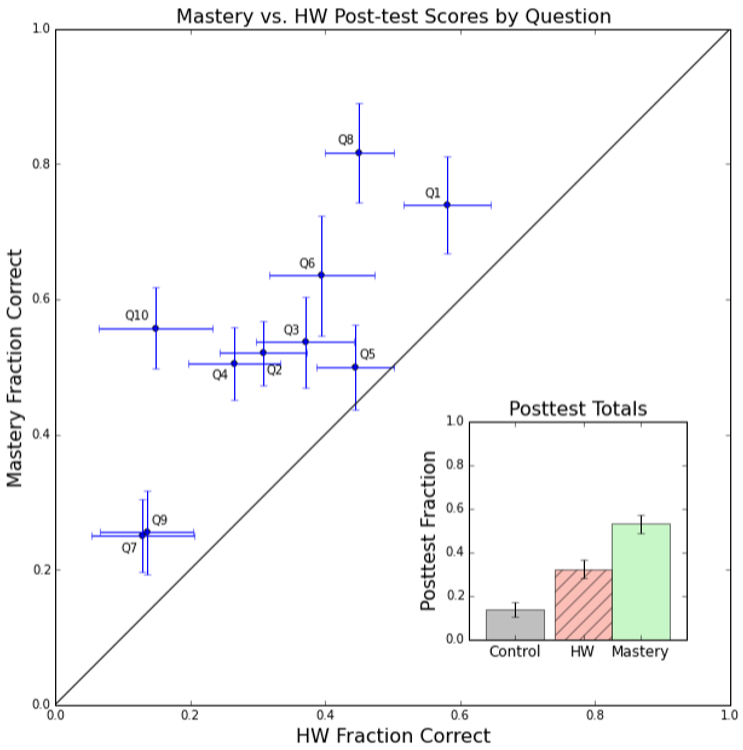


Figure 4.6: Outset: Mastery group students' average score compared to homework group students' score by question. Points above the line indicate that the mastery group outperformed the homework group. Inset: Posttest average scores for both matsery and homework groups, with control average score (no treatment) from the previous experiment included for reference.

from a potential difference in a planar geometry of two charged sheets to find the charge density of one unknown sheet. This required a sophisticated synthesis and transfer from the training materials.

4.4 Discussion

The experiment was designed to evaluate a mastery delivery system for learning materials compared to more traditional online homework. The students in the mastery group saw multiple versions of questions, as needed, coupled with delayed feedback for each iteration and followed with narrated animated solution videos, while students in the traditional homework group had unlimited tries on a single version, immediate feedback by question, and access to notes, colleagues, and online resources. On the written posttest, the mastery group signficantly outperformed the homework group, despite spending nearly identical time on the training exercises. They performed better on the near transfer question, and the questions which required synthesis of the skills during the treatment. The mastery delivery system takes advantage of formative assessment, multimedia principles, the worked example effect, and other learning principles which may account for students' improved performance.

The students in the traditional homework group exhibited behaviors they believed would get them through the treatment as quickly as possible, often resorting to trial and error to find the answer. Because the

stakes were extremely low for submitting an incorrect answer, there was little incentive to check their work and students occasionally used incremental guessing to attempt to find the correct answer. The homework students had access to notes and online resources but rarely used them until they had exhausted their initial ideas or guesses. This may explain why students in the traditional homework group submitted significantly more attempts at the answers. In short succession of each other, they could transfer one question's strategy to a similar question, but failed to do the same when the questions were spaced out and on different levels. In contrast, the mastery group worked more conscientiously in response to the higher stakes, double checking their work and using sense-making strategies to avoid the penalty of redoing an entire set of problems, observed anecdotally during the sessions. The students showed visible signs of stress when submitting answers to be graded, and likewise were vocal about disappointment or excitement after their sets were scored. The mastery students showed outward signs of investment in the mastery activities. This suggests that the higher stakes of the submissions increased student engagement and sense-making, which led to a larger learning gain from as the levels progressed, and higher performance on the posttest.

The changes to the experiment from the previous clinical trial had negative and positive consequences with respect to students' frustration. Breaking the larger levels into smaller pieces gave students more opportunities to feel success, which they often celebrated aloud, but the increase of the mastery threshold from 85% to 100% slowed some students down on the final levels. We expected that breaking the material into smaller masterable chunks would help students work through the material faster, but both experiments had students averaging around 2 hours on the treatment materials. In the last levels, students were expected to keep track of many small steps and negative or positive signs, so giving them no leeway for error made it more difficult to master those problems. Some students complained of fatigue and frustration, as in the previous experiment. This is a consequence, in part, of the requirement that students complete the whole assignment in a single sitting, which would not necessarily be true for students doing homework for a course.

The purpose of this clinical trial, in part, was to see the suitability of mastery style online exercises for a course, as an improvement over the traditional online homework used in courses. The previous and current clinical trials gave evidence that the mastery delivery system was more effective, and the narrated animated worked solutions are effective correctives to use within the mastery delivery system. Students are able to master the materials with the support of the worked solution videos, as well as better carry their learning into near and far transfer problems. The results from these mastery clinical trials were encouraging and set the stage for its inclusion in an introductory course.

Chapter 5

First Semester-Long Implementation in Physics 100

5.1 Introduction

Encouraged by the local positive clinical results in favor of mastery learning, the next step of implementation was a course-wide homework reform. Physics 100: *Thinking About Physics* was a natural place to try mastery homework for many reasons.[1] Recall that the course is a preparatory course for engineering students who may be underprepared for the calculus based introductory physics sequence.[2] The mastery literature asserts that one of mastery's strengths is its ability to help students at all levels, and for a course that intends to bring students from different educational backgrounds up to the same target readiness, mastery learning is well suited. Because the course is outside the main physics course sequence, there is more flexibility for experimental methods; different attempts to improve retention of at-risk students have been used in the course before. Lastly, from a practical standpoint, the pace of Physics 100 is slower than other courses; there are only eight weeks of material spread over the semester. To implement mastery-style homework requires mass amounts of content creation, so tempering the workload by using a class that covers less material also made sense, especially for a first attempt at a course-wide renovation.

It is worth noting that a clinical implementation is very different than an in situ course implementation. In our clinical trials, students were volunteers who had intentionally set time aside to participate in the activities and had already completed the introductory mechanics course that Physics 100 prepares students to take; the students in Physics 100 are more novice and also dealing with the transition from high school to college, most of them first-term freshmen. The diversity in experience and expertise of the students in Physics 100 was a reason that the course was an appealing choice, but most students who participated in the clinical trials had successfully passed at least one introductory physics course, which washes out some of

[1] Parts of this chapter were part of the published article, "Mastery-style homework exercises in introductory physics courses: Implementation matters" in 2018 [11]. Some of this content also appears in Noah Schroeder's thesis, "Mastery-Style Exercises in Physics" [65].

[2] More information about the course and its population are given in the Introduction, Chapter 1. Background on mastery literature is documented in Chapter 2.

the differences between students, either by acting as a barrier to students who are not prepared or providing them additional experience that made their previous differences less prominent. The implementation into Physics 100 required more care and effort than the clinical trials and had many more uncontrolled variables. This chapter documents some successes and struggles of students in the first year of mastery homework's implementation, many of which were unanticipated, and lays the groundwork for improvements in its next iteration.

5.2 Methods

Our implementation for mastery homework used the online platform smartPhysics (rebranded FlipItPhysics) to deliver tests and re-tests, as well as correctives in the form of narrated animated solutions, the same format used in the clinical studies. For each of the eight weeks, we developed specific competencies for different, defined units, which were usually successively more difficult and building off each other, targeting the skills we hoped students to develop that week. The decisions regarding what these skills should be were discussed, informed by emphases placed on past exams, student performance on those exams, and anecdotal experience of student weaknesses after teaching the course for many semesters. From those competencies, we developed problems to target the learning goals prescribed, trying again to create many similar (but not identical) versions for each set. On average, there were usually three or four levels in each week, with four similar versions of each unit, around five questions long. In total, we created over 500 questions and solution videos which were coded and put into the homework system, delivered over eight weeks to about 500 students. A list of levels and associated competencies are included in Appendix C. The threshold for mastery was a perfect, 100% performance. To prevent frustration from rounding errors or careless mistakes, we created the questions in multiple choice format, but also intentionally included attractive conceptual distractors as answer choices. These mastery exercises made up nearly all of their regular homework, which was completed online weekly and was implemented in Physics 100 for the first time in Fall 2014.

Versions of the same level relied on the same principles but were intentionally created to ensure that the algebraic solution to the problems were different on each version, to keep students from transferring the solution from the previous version directly to the new problem without thinking about principles. Students were randomly assigned to a first version (versions labelled 1-4) and then worked consecutively. This was different than the clinical trial, where each student saw the same version to start. This was done to reduce the possibility of copying between students and to give us a sense of the difficulty of levels relative to each other, since each version then had many students attempting it as their first version. Within each version, there was also random seeding of numbers, if the problems were numeric; even two students receiving version 3, for example, would have different numerical answers, even if they had the same algebraic form of the answer. If students exhausted all versions, they would see another iteration of a previous version, but with different numbers. For example, a student who began on Version 3 would next see Version 4, then 1, 2, and 3 again. They were given as many tries as needed to achieve mastery, and were restricted from later levels until they scored 100% on one of the versions. After grading a version, students had access to the narrated worked solution videos for all problems, regardless of correctness on each question. When students called their next version (if they had not mastered), the video solutions for the previous version were no longer available. If moving to a new level, the videos remained available for the previous mastered versions of each level. A schematic for this delivery system is shown in Figure 5.1. Because the homework assignments were weekly,

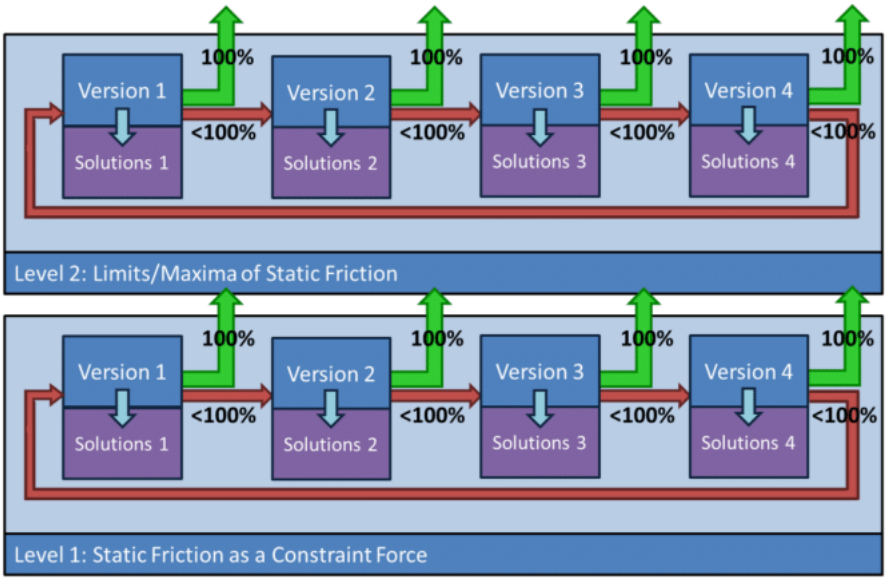


Figure 5.1: A schematic of mastery-style exercise format, including multiple versions of each level (which contains a number of individual questions). Each assignment typically consisted of a few levels, which open upon mastery of the previous level. Solution videos are available to students after grading their current version.

students could work at their own pace, but were not entirely free of time constraints, as recommended by Kulik et al [38].

On their first day of discussion (called a recitation section at many universities, which is before their first lecture in this course), students were introduced to the mastery homework system by their Teaching Assistants (TAs) and given a simple mastery assignment to practice on. The first discussion section is generally spent going over the course structure and onboarding students to its objectives and logistics. The course runs on a recurring cycle for each unit of content, which is shown in Table 5.1. The unit generally opens with students becoming familiar with new material through online prelectures (narrated animated videos) paired with checkpoint questions (to gauge student understanding of material prior to lecture). Students then attend lecture on Fridays, work on homework pertaining to the content which is due on Tuesdays (and have the potential to attend office hours on Monday, prior to its submission). Discussion section typically is an opportunity for students to clear up misunderstandings and work more in depth on the content with their peers and TA in a smaller class setting, which has sections on Tuesday through Thursday. Thursday evening, students must submit their online quiz, to act as a final assessment of that material before the cycle begins again. The online quizzes are given as delayed feedback, meaning the students receive feedback on correctness only after the deadline, to simulate a typical quiz. The course uses this cycle for the first 10 weeks, which covers 8 weeks of content; the end of the first week starts the cycle and students have one week off from homework for their midterm exam. The rest of the semester focuses on reviewing concepts and focuses on problem solving strategies. In those weeks, prelectures, lectures, quizzes and discussion sections remain, but there is no homework. Students take both a midterm exam and final exam.

Data collected during the semester was primarily through the online homework system. Every student interaction with the system created an entry for its time and nature, so we were able to see what students were answering, when they prompted video solutions and called new versions. Using the differences between

Table 5.1: Activities and their timing for a typical unit of content in Physics 100.

Day	**Activity**
Before Friday	Prelecture and Checkpoint
Friday	Lecture
Monday	Office Hours
Tuesday, 8am	Homework Due
Tuesday	Discussion Sections
Wednesday	
Thursday	
Thursday, 8pm	Online Quiz due

time stamps, we estimated time spent on different activities, keeping in mind that students had the ability to leave an assignment and return later; if a student spent more than ten minutes between interactions, it was assumed that they had disengaged and time was not added to their cumulative time. The course also included an affect survey given in Weeks 1 and 5, which asked students to self-report their experience with the mastery homework, on a Likert scale from Very Frustrating to Very Encouraging and gave space for open-ended responses. This implementation took place in Fall 2014, and the online quizzes and two exams were repeated from Fall 2012 to be used as another assessment of changes in learning. Historically, students' online quiz scores have best discriminated and predicted students' success on exams and in the course; a check between Fall 2012 and Fall 2013 on identical online quizzes showed that students were not significantly different from one semester to the next, which is plotted in Figure 5.2. The student populations from semester to semester are not significantly different and we expected the same to be true of Fall 2014, which allowed us to confidently make those comparisons.

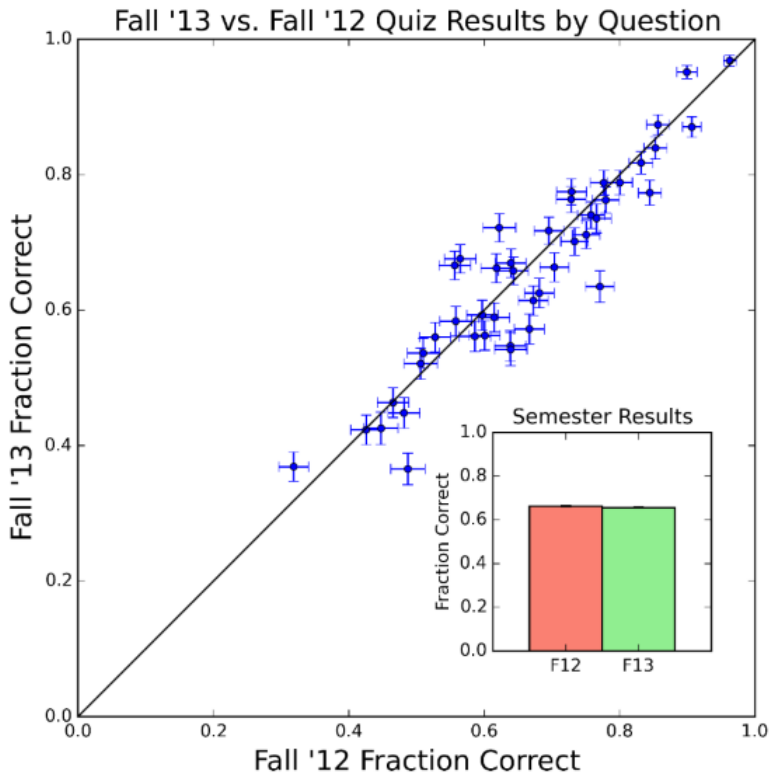


Figure 5.2: Outset: Scatter plot of students' success rate on each online quiz question in Fall 2013 versus Fall 2012. Inset: Average online quiz scores summed over each semester.

5.3 Results

5.3.1 Performance on Quizzes and Exams

Assuming that the students in Fall 2014 were an unremarkable population compared to previous years', we compared students' scores from their quizzes and exams to scores from a non-mastery year as a measure of learning. The quizzes, midterm and final exam were repeated from Fall 2012, so this was the natural population to compare to. Figure 5.3 shows students' quiz scores by question in Fall 2014 compared to scores in Fall 2012, with average scores on the inset bar plot. The scores showed a small but statistically significant improvement when averaged over the semester, $70.5 \pm 0.3\%$ compared to $66.3 \pm 0.3\%$, which was an effect size of 0.6 with $p < 0.0001$. Likewise, a comparison of student performance on the midterm is shown in Figure 5.4. The midterm was given after 7 weeks (of 8) of the homework, and was identical for both groups. There were some questions which showed statistically significant improvement, but overall the scores were statistically similar and showed no improvement.

After such positive results from the clinical trials, these modest improvements were surprising. However, closer look at students' interactions unveiled many differences between the behavior of the students in our semester-long course and the behavior of the students in the clinical trials.

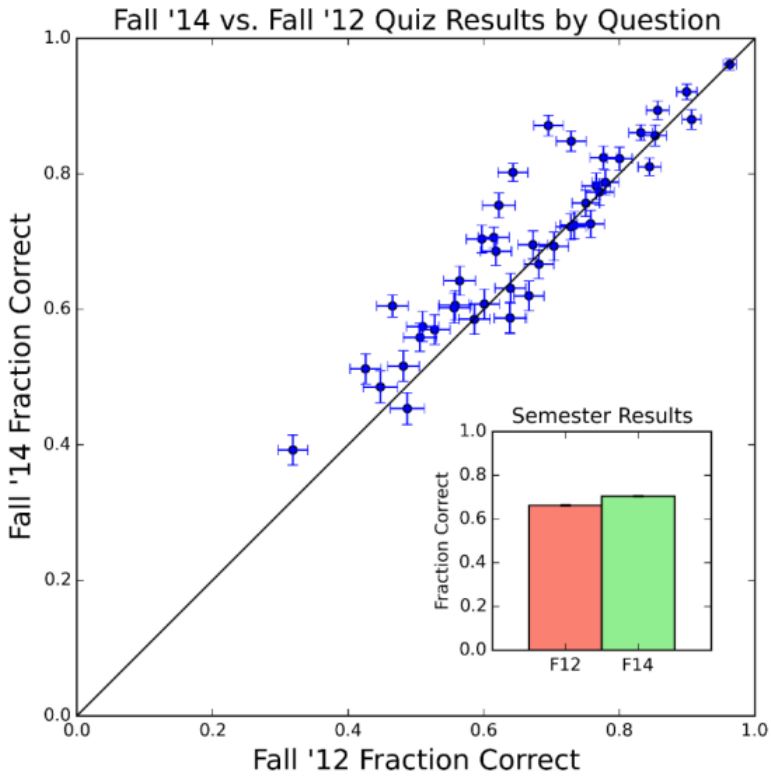


Figure 5.3: Outset: Scatter plot of students' success rate on each online quiz question in Fall 2014 (mastery) versus Fall 2012 (no mastery). Inset: Average online quiz scores summed over each semester.

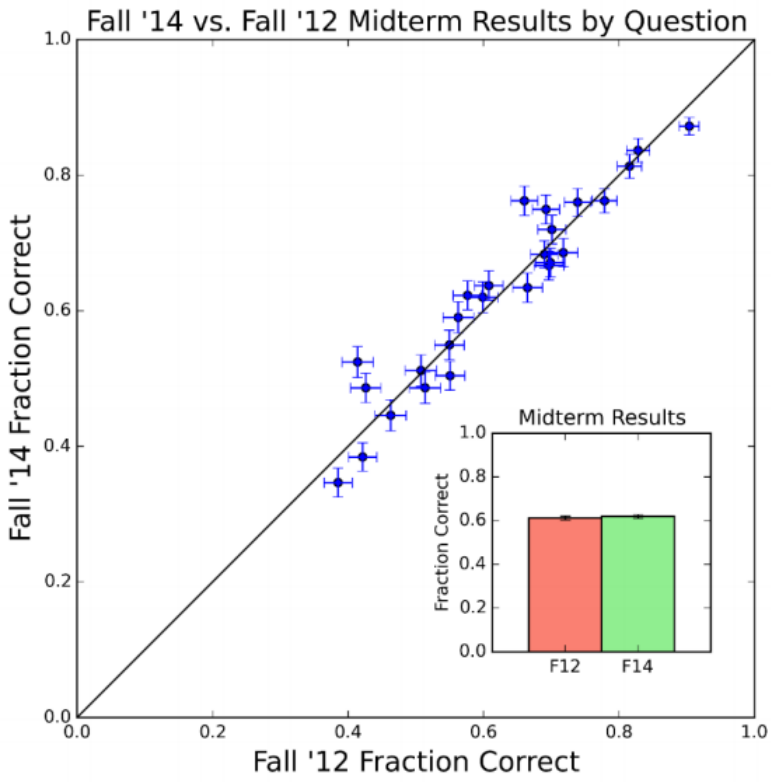


Figure 5.4: Outset: Scatter plot of students' average scores on each midterm question in Fall 2014 (mastery) versus Fall 2012 (no mastery). Inset: Average midterm score for each semester.

5.3.2 Student Engagement

Some of the most striking results from the semester spoke to student attitudes and behavior; it is not surprising that first-term students over an entire semester would engage differently than volunteers for a single clinical trial, but the gravity of that difference was unexpected. Most students began the term using the homework as it was designed, but their useful behaviors deteriorated over the semester.

One dramatic behavior that emerged was students' tendency to "skip" through levels, spending very little time answering the questions. We observed that as the term went on and the difficulty of the levels increased, many students were actually spending less time answering all the questions in a level in a given try. We believe that this behavior arose from the use of one of two nonproductive strategies or some combination. One strategy would be to try to take advantage of the multiple-choice format by simply choosing answers at random to questions that they had no idea how to answer. Statistically, this strategy is very unlikely to be successful, as most levels had around 5 questions and most had at least 5 answer choices. Another strategy was to take advantage of the cyclic nature of the finite number of versions. Since a given level had four unique versions which cycle, students could study the solutions to the first version they were given, then choose answers quickly for next three versions to return and apply what they had learned from the first version's solutions to the repeated version (albeit with different numbers) on their fifth try. Anecdotally, many students readily admitted to this strategy.

A signature of the second strategy is seen in students' pass rates as a function of their attempt. Figure 5.5 shows the fraction of students who mastered, according to their attempt number for three levels at different times in the semester. The first and fifth attempts are the same version, and the repeated version is indicated in orange. In the first level of Week 2, students worked diligently through all attempts, as seen by a somewhat steady rate of mastering for each try. As the semester progressed, students' mastery rates on intervening tries dropped. The last level of Week 3 shows students working earnestly on their first attempt, lower rates of mastery on tries 2-4, and then a bump again at the fifth try. By Week 7, mastery rates were low on all attempts but their fifth.

If material is particularly difficult, the spike in performance at the repeated versions could be natural; students working within the system as it is intended would still have an advantage on a version they had seen previously. By looking at the distribution of time spent by students on their last attempt before the repeated version on these same levels, plotted in Figure 5.6, we likewise see that the number of students spending less than one minute on a version increases as the semester progresses. By Week 7, over half of the students who do not master on their fourth attempt were spending under a minute on a version.

Looking at the average amount of time students spent per try on their second through fourth tries across all levels in the semester, we identified students engaging in these "skipping" strategies as those who spent less than one minute on an attempt (or less than 1/4 of the median pass time, if that is under a minute). The trend over the semester is shown in Figure 5.7. As the semester progressed, the number of students spending very little time with the levels themselves increased to nearly half by the end of the semester. These fractions are not of the students who had not yet mastered, as in previous plots, but include all students who attempted each level.

Despite low mastery rates in many levels, students' interactions with the solution videos were also less frequent than for students in the clinical trials. Broken down by try, Figure 5.8 shows the average fraction

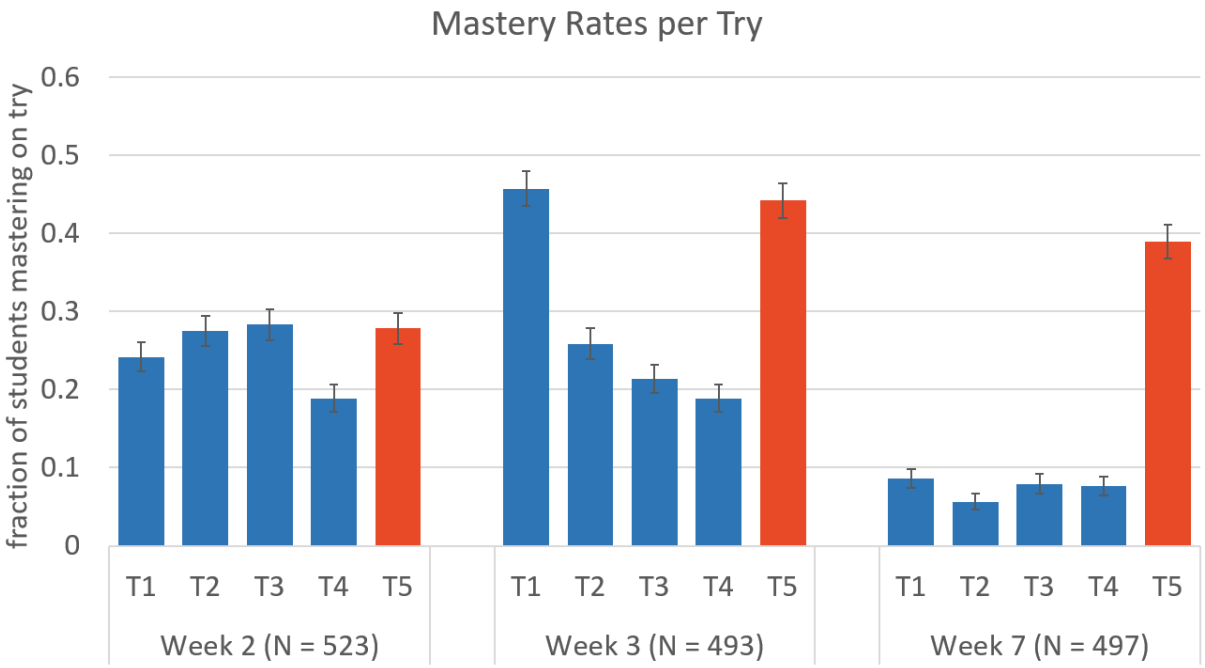


Figure 5.5: Fraction of students mastering by Try (T1 is Try 1, etc.) for three levels in different weeks in the semester. The orange versions are repeated, and students' performance is much higher on repeated levels, which is more prominent as the semester progresses. It should be noted that the fraction for each try is the fraction of remaining students, meaning that each subsequent group is smaller and a selectively contains students who struggled on previous attempts.

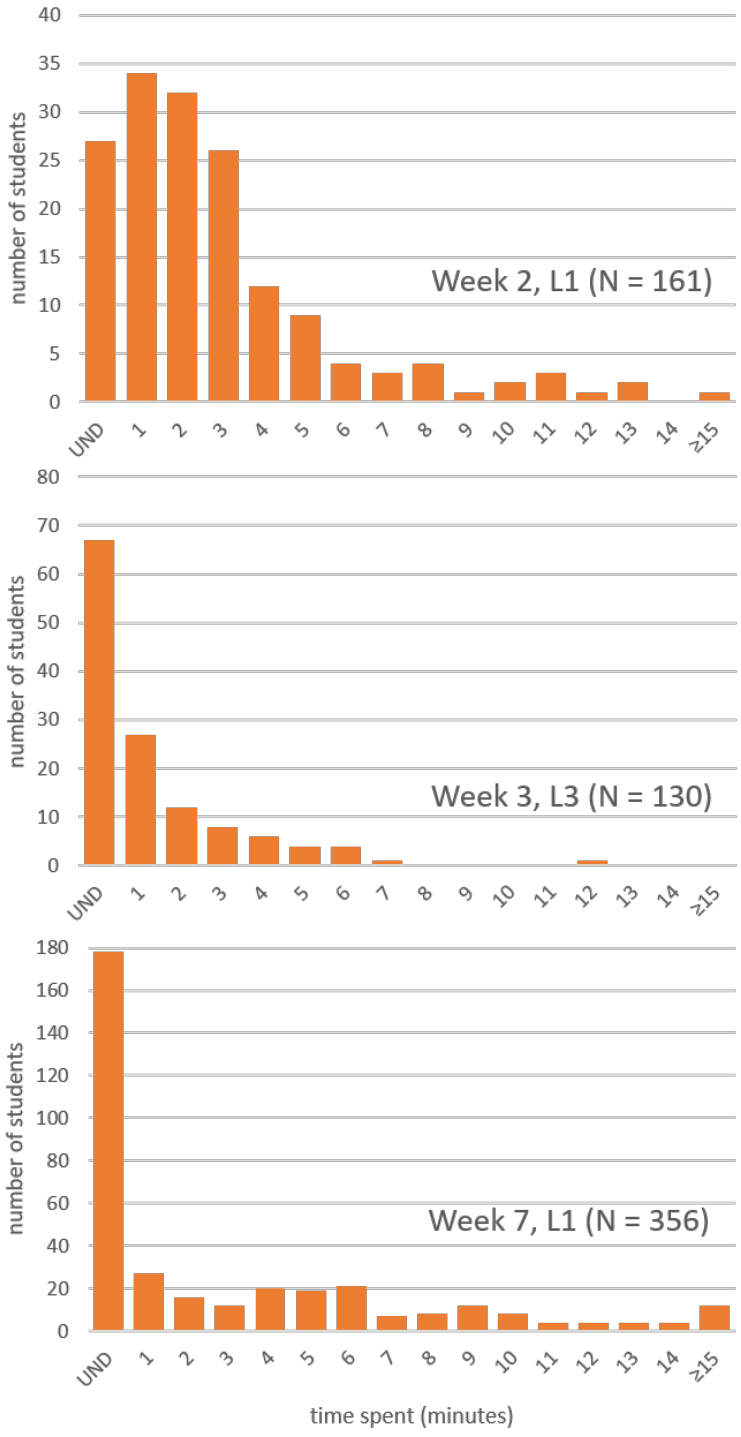


Figure 5.6: Histogram of time spent on students' fourth attempt, if they did not master on that attempt. Earlier weeks show more spread in the distribution, while later weeks show large numbers of students spending less than one minute on their attempt of multiple questions. Note that the scale of the vertical axis is different in each plot, with larger numbers in later weeks.

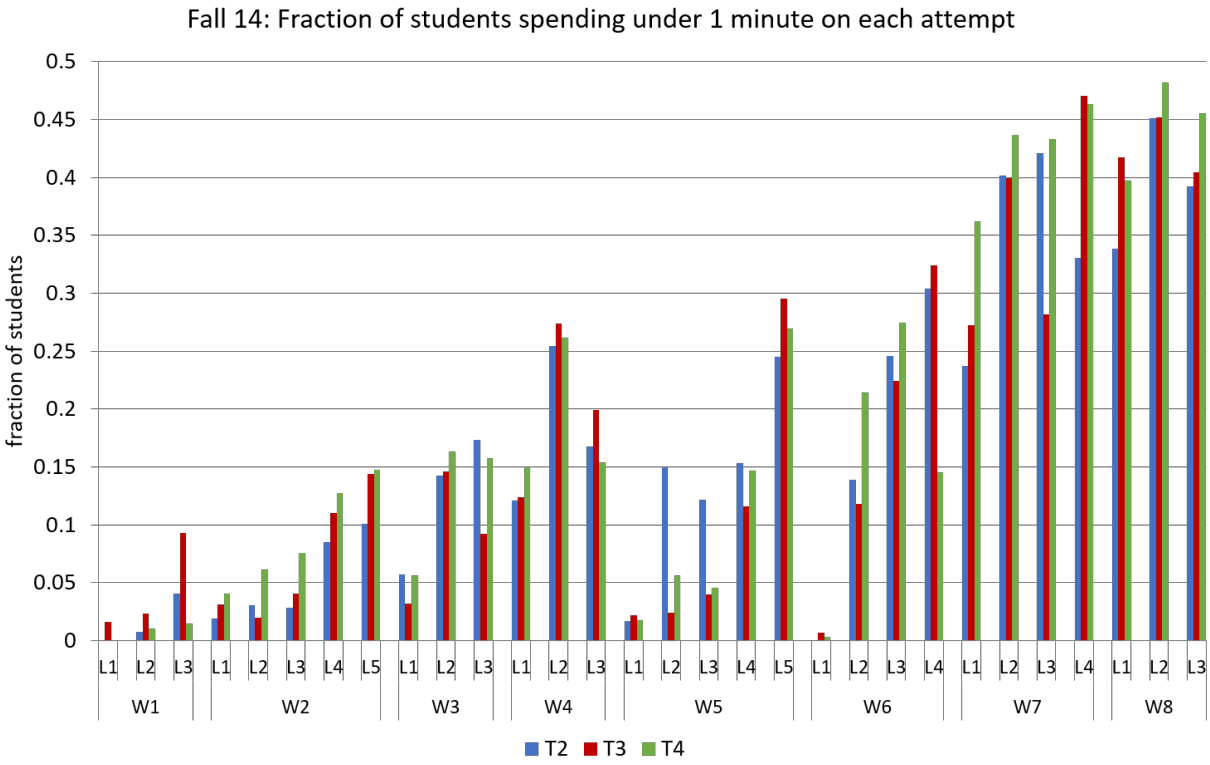


Figure 5.7: Fraction of students spending under one minute, or under 1/4 of the median pass time (whichever is lower), on an attempt for each try, Try 2 through Try 4 (T2-T4) at each level throughout the semester.

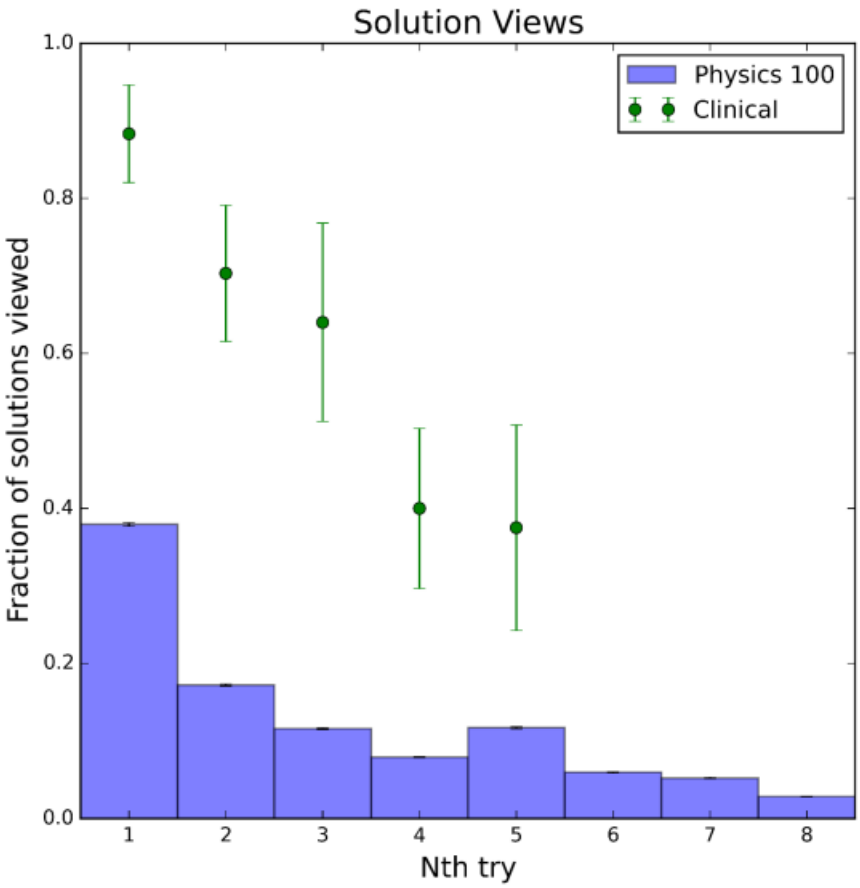


Figure 5.8: Fraction of solutions viewed by students in Physics 100 in Fall 2014 compared to students in the second clinical trial. Viewing is defined to be watching at least 20% of the video's length, and the fraction is out of the total number of incorrect questions. Using this definition, it is possible to get a fraction larger than 1, if students watched videos to correctly answered questions.

of solution videos watched by students in Physics 100 compared to students in the second clinical trial. The threshold used for counting a video as "watched" was set to 20% of the length of the video, and the fraction was out of the number of problems the students answered incorrectly. Both instances had a diminishing fraction of solutions watched as the number of tries increased, but the Physics 100 students began at a much lower fraction, less than half of the rate of the students in the clinical trial. Checking their solution watching over the semester also reveals that productive behaviors declined with time. Using a 10% threshold for watching, Figure 5.9 shows the students' fraction of solution videos watched week by week.

While we did not expect students in the semester-long course to behave as diligently as students in our clinical trials, we did not anticipate such dramatic differences, nor the decline over the semester. These behaviors can in part be explained by declining student affect, described in the next section.

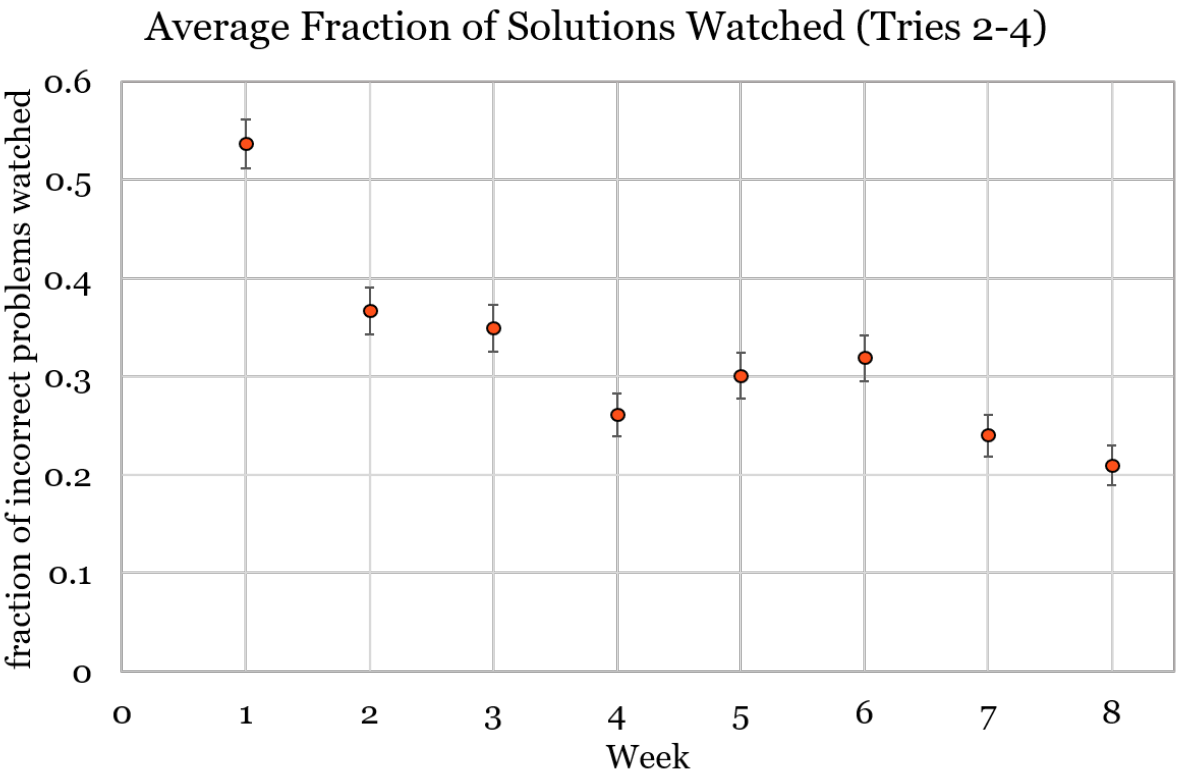


Figure 5.9: Average solution viewing fraction over the semester. Threshold for viewing was 10% of the solution video's length, and the fraction is out of the total number of incorrect questions. Using this definition, it is possible to get a fraction larger than 1, if students watched videos to correctly answered questions.

5.3.3 Student Frustration

Over the semester, students' affective response to the mastery homework decreased in tandem to the decline in productive behaviors, perhaps motivating their unproductive engagement. Students completed the same survey in Weeks 1 and 5, which asked them to comment on the mastery homework. In response to the prompt, "Which statement best reflects your experience working through the mastery homework?," students answered on a 5 point Likert scale from "Very Encouraging" to "Very Frustrating." Their responses in both weeks are shown in Figure 5.10, showing a shift in attitudes from the beginning to the middle of the course. In Week 1, around 30% of students self-identified as being frustrated, which doubled to approximately 60% by the fifth week.

Students were also given the open-ended prompt, "Please write any comments that you may have about the mastery homework in the space below," where they could share more personalized responses. Like their multiple choice survey responses, there was a shift from the types of feedback received in the first iteration of the survey to the second. In Week 1, many students appreciated the mastery homework and commented on its ability to help them understand the material. Some selected quotations are below:

> "I like the mastery homework because I have a full understanding of the material before advancing to harder material."
>
> "I learned well from having to get 100% correct before moving on and it helps that the questions are similar but are still different so I really had to learn the material to be able to pass through."
>
> "I like how new problems are generated, so you for sure learn to do the problems, not just guess."

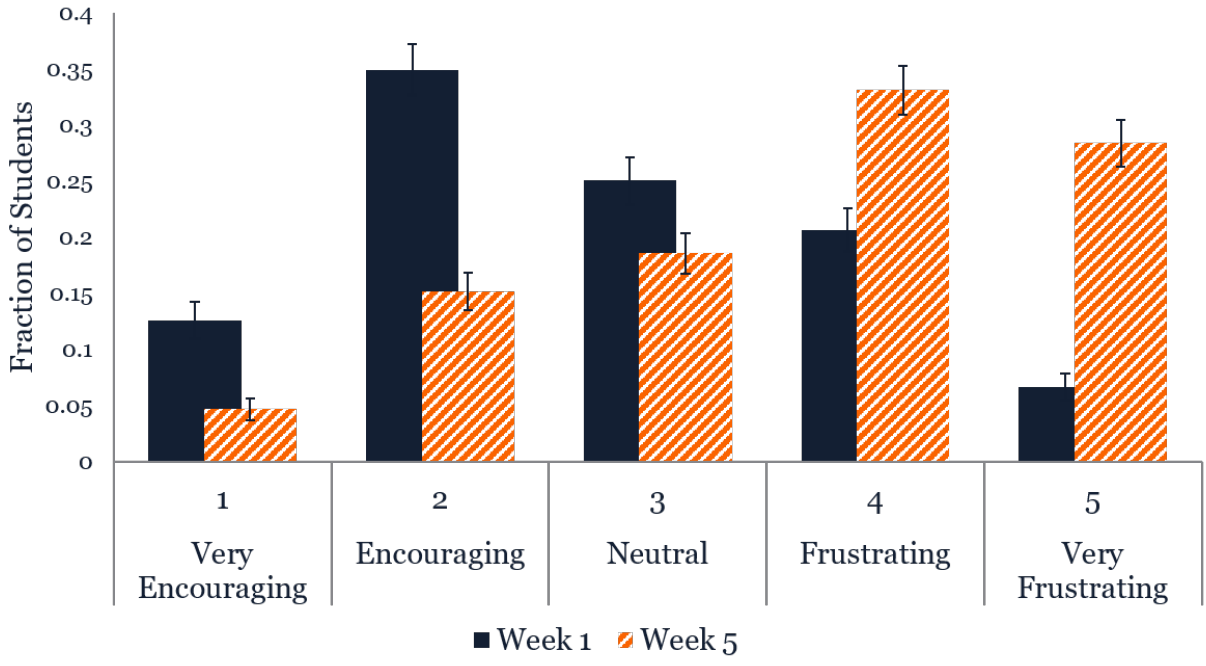


Figure 5.10: Students' self-reported responses to the question, "Which statement best reflects your experience working through the mastery homework?" on a scale from "Very Encouraging" to "Very Frustrating" in Weeks 1 and 5.

"For as many times as i may fail, being able to keep working at new problems until I fully grasp the concept is very encouraging. I feel that I have already learned so much just from the mastery homework."

Many student responses were positive, but some students also mentioned their frustration in redoing sets of problems after missing a single question, as seen in this response from Week 1:

> "It's a bit annoying that you have to repeat the entire set of questions if you only missed one, though."

By Week 5, responses were less positive. The frustration in repeating problems was mentioned much more often, cited as unfair or unnecessary, despite students' praise of the repetition earlier in the term. The same student who gave the extremely positive review of mastery in italics above gave the curt italicized quotation below, included with their colleagues' responses:

"I don't like that if you get one problem wrong you have to redo the entire thing."

"The mastery homework that has more than 4 questions that you have to get right is extremely frustrating as getting even one wrong makes you have to redo the problem entirely."

"not fair to have to redo the entire section if I just missed one"

"Requiring a student to redo a set of problems because they missed one out of six is unnecessary."

5.3.4 Potential Causes of Frustration

As seen above, one cause of student frustration was a lack of agency in how often to repeat practice. Although repetition is inherent in mastery learning, students also cited that the versions were distinct from each other, which amplified frustration. The versions were designed to call on the same principles and we believed they were isomorphic, but students performed differently on versions within the level, revealing differences to students we had not anticipated. The mismatch in difficulty between versions was also cited as a deterrent for studying solution videos; one student compared the experience to being a mouse in a maze, where they were taught to solve it and then put back at the beginning of an entirely different maze. Many students did not believe that the solution video to one version's questions would help them navigate the next version's.

Because students were randomly assigned to their first version within a level, data allowed us to compare their scores on the first attempt as a proxy for the difficulty of versions and questions within each level. A severe example of mismatched levels is shown in Figure 5.11. This level required students to synthesize skills from the previous levels to solve a Newton's Second Law problem, but Versions 1 and 3 involved rotated coordinate axes, a problem on a ramp, while Versions 2 and 4 used traditional coordinate axes. The stark difference in performance on these versions gave us information about student difficulties that we had not anticipated.

Considering the mismatch of versions over the entire semester, a threshold was set for 20% performance difference to constitute a level as being mismatched in versions. This could either be overall mastery rate (one version had at least a 20% difference in mastery rate than another version) or identified by having at

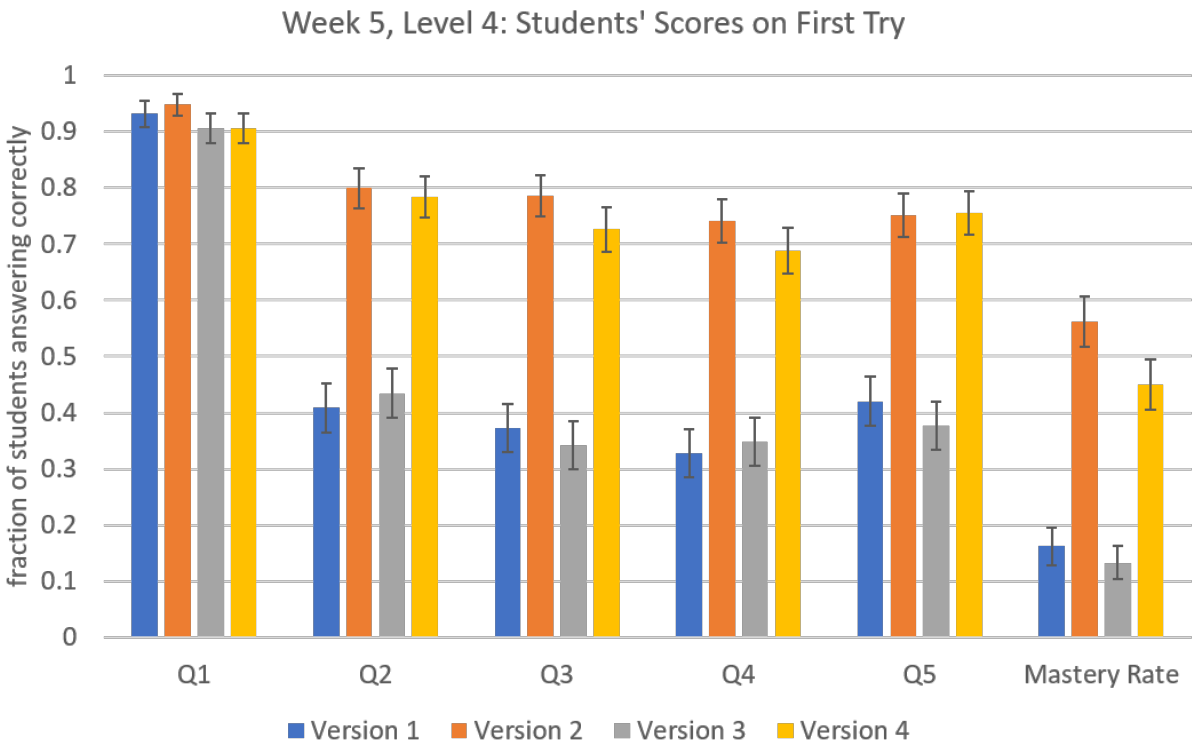


Figure 5.11: Students' scores by question and mastery rate for their version on their first attempt, which can be used as a proxy for version difficulty.

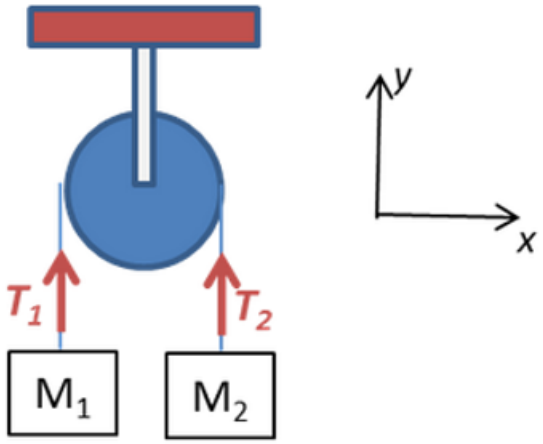


Figure 5.12: An example of the setup for a two-body Newton's Second Law level. Students were asked to correctly identify equations for each block to eventually solve for their acceleration.

least two questions which had a 20% difference in performance between versions. Some levels had smaller differences between questions which cumulated to a systematic difference over the whole set. Others had specific questions that were dramatically different, but a floor effect in the mastery rate made identifying them by mastery rate alone ineffective. In the level shown in Figure 5.11, both conditions were met; there were specific questions showing large differences which translated to a difference in the overall mastery rate. Identifying levels which satisfied either one or both of the conditions, 20 of the 32 levels were mismatched by our definition. Their distribution throughout the semester was consistent, with fluctuations but no general trend; each week had at least one mismatched level.

The data collected from the semester allowed us to identify skills which students struggled with that were taken for granted in the creation of the content. In some cases, these were related to math skills, such as the example discussed with the mismatched versions. We expected students to be proficient in rotating vectors to new coordinates, but the students needed support. Other examples of math difficulties were also revealed through the data. Another notable example was students' performance on a two-body Newton's Second Law level, one which did not have mismatched versions. An example of one of the version's situation is shown in Figure 5.12; students were asked in questions 1-4 to (1) identify the correct equation of motion for M_1, (2) identify the correct equation of motion for M_2, (3) specify that the accelerations were the same magnitude but in different directions ($a_1 = -a_2$), as specified by the coordinates, and (4) specify that the tensions on each block were the same magnitude. Of the students who could correctly answer questions 1-4, only 47% could correctly solve for acceleration. Despite having the correct equations and relationships necessary to solve the problem, students struggled to perform the algebra to solve the system of equations. Most of the content in the course was not in support of math competencies, but relied on students' ability to complete math tasks.

5.4 Discussion

The primary takeaway from the first year of implementing mastery-style homework was that if we want the students to use the materials productively, affect needed to be addressed directly. Because the results from our clinical trial had no such issues, we neglected to consider students' emotional response to the system while dealing with it for an entire semester. Many students were not using the homework as intended, but rather circumventing the learning portions (sincere attempts at the problems and the solutions) in order to get their points as quickly as possible. Though this behavior is not completely unexpected for first-semester college students, the scale and impact of student frustration and desperation was greater than anticipated. As the term progressed, these issues compounded; students became frustrated, used the system in ways that did not help them learn, and then felt more frustrated when they could not demonstrate learning. To address this, we needed to look for ways to improve student attitudes about mastery-style home-work to encourage them to use the system in a useful way.

A potential cause of frustration was mismatch among versions of the same level, as far as content and difficulty. Wording or coincidence among multiple-choice answers sometimes made specific questions harder or easier than in other versions, which led to either frustration by passing students before they were prepared for the following level or frustration because the practice and solution videos did not prepare them for an unusually difficult problem on their next version. Further, scenarios that we considered equivalent had vast performance differences. One reason that students cited for their lack of interest in the solution videos was that they felt that each new version was unrelated to the last, and thus had little motivation to review missed questions. Literature has shown that novices and experts tend to classify problems differently; experts are more likely to classify problems as similar based on underlying principles whereas novices tend to focus on surface features [108]. This may explain some of the students' belief that versions were not equivalent, but their different performance rates on versions suggests that as "experts," we underestimated how details affected the complexity and skills necessary to solve problems in different scenarios.

Students' response to the mastery homework may have been informed by their goal orientation, as described by Dweck [34]; mastery is useful for students with learning goals orientation, who are motivated primarily by wanting to gain competence in skills. The language and behaviors by students suggest that they were primarily concerned with desperation for points and frustration with repetition, more often associated with performance goal orientations. Especially because the scope of the audience of Physics 100 extends to all of the engineering college, it is likely that there are students in the course who do not have an explicit interest in learning physics, but are taking the course to fulfill that requirement. The benefits of mastery would then be diluted for students approaching the exercises with performance rather than learning as their motivation.

Moving forward, it was clear that affect needed to also be addressed directly. General student attitudes towards the homework were resentful, which may have also been due to students'perceived lack of agency. The language used by students portrayed the situation as something "done to them" as opposed to something "for them," which did not promote students to take ownership of their learning. Rather than a tool, the homework was seen as an obstacle to overcome by any means. Our main objective for the following year was to reduce frustration and shift student attitudes to improve student morale and increase productive behavior.

Chapter 6

Second Semester-Long Implementation in Physics 100

6.1 Introduction

The first implementation of mastery homework in a semester-long course unveiled areas for improvement, with an overarching need to address student frustration. Attention to student morale was the priority for its next implementation, which was addressed via content reorganization to support student weaknesses (such as some math skills), altering the homework's delivery method, and direct discussion of the role of homework, the final two intended to foster a greater sense of ownership in the students.[1]

In education literature, there is significant research emphasizing the importance of productive struggle, particularly in mathematics [109–122]. Engaging in productive struggle has the potential to activate prior knowledge as well as intuitive ideas [118, 121], and has been shown to help students identify gaps in their understanding [122, 123]. For areas of study where struggle is anticipated, framing struggle as a productive process can improve students' persistence and positively impacts their performance [124]. Many of these references emphasize the importance of scaffolding to keep students' struggle productive, rather than dissolving into unproductive frustration. One recommends differentiated instruction [112], a type of personalized instruction which allows for flexible content, delivery, assessment, and styles of learning; mastery learning provides flexible content depending on student readiness, but still treats all students to the same types of assessment and content delivery. Providing a lesson to reconcile and consolidate after a student failure was also shown to be more effective than student failure on its own [118, 121], which mastery learning does effectively. Keeping students' engagement in a productive struggle domain requires problems to be challenging but also "within reach" [110], alluding to Vygotsky's Zone of Proximal Development (ZPD) [32].

A conference paper by McQuiggan et al documents efforts through an online intelligent tutoring system to predict and respond to student frustration using affect recognition models that code for the amount of time students spend on different activities, which areas of the learning environment they are spending most of

[1] Parts of this chapter were part of the published article, "Mastery-style homework exercises in introductory physics courses: Implementation matters" in 2018 [11].

their time, and physiological data (via a physical biofeedback apparatus) [125]. While this type of predictive model is beyond the scope of what is capable for our system, their suggestions to give feedback on students' use of the activities, rather than just correctness, and providing extra scaffolding for struggling students are relevant. Essentially, giving students feedback on their holistic use of the system and responding in time was successful at mitigating potential frustration.

As part of scaffolding learning, the homework's first iteration provided clues as to which subjects were not "within reach," and surprisingly, students often struggled with math competencies that were required but not explicitly taught. Students entering Physics 100 historically have about a 30 on the ACT Math score (of 36), which led us to believe that these competencies could be taken for granted. Cabellero et al describe a framework for understanding synthesis and application of math knowledge in the context of upper-level physics courses; the ACER framework describes four components: Activation of math tools, Construction of the model, Execution of mathematics, and Reflection on the results [126]. Their motivation was also to discover why students who were proficient in math class struggled to apply their math skills to physics problems. They found that often difficulties were in the Activation step and synthesizing their math with the physical model [127]. Students often did not reflect on their answers unless prompted, which could also account for student errors. In our case, our students are typically good at math, which suggests that recognizing and activating the correct math competencies may be the issue.

Another theme that emerged which contributed to student frustration was a lack of students' sense of ownership over their learning on their homework. Some models which describe student ownership typically include aspects of "students' level of autonomy, choice, control, interest, investment, or responsibility with respect to the purpose, design, implementation, or assessment of an educational activity" [128]. In the context of our homework system, we intended students to have autonomy in practice by having control over movement through the material, but the penalty for not completing the activities (receiving no points) removed students' sense of choice to engage, and rather created an obstacle for them to overcome. This is consistent with Podelefsky's explanation of the analogy of 'estranged' learning to estranged labor [129], which they drew from Lave and McDermott [130]:

> "when learners lose ownership over the means of production of knowledge, they become alienated from both the product and process of learning. Further, they enter into a relationship that is competitive, rather than cooperative, between themselves, teachers, and other participants in the school system."

This well describes the response we saw to the mastery homework; students treated the homework as an oppositional force to beat, rather than a tool designed to help them. There are multiple reasons for instructors to care about their students' sense of ownership, such as their sense of pride [131] or persistence through the sciences [132], and as seen directly in this work, student motivation [133–135]. Besides the obvious benefits of students as people feeling agency in their lives, which extends to their academic domain, a lack of ownership has negative effects on student engagement, which we saw hurt their potential to learn.

Moving forward with the second iteration of mastery-style homework, keeping in mind students' sense of ownership and attending to difficult concepts via more supporting scaffolding were imperative to reduce frustration. We hoped that by addressing student frustration, we could move students into a domain of *productive* struggle, rather than debilitating frustration, and that improved learning, behavior and attitudes would follow.

6.2 Methods

The effort to improve student engagement began with a change to the homework's delivery method. To quell students' tendency for skipping intermediate versions or resorting to guessing, students were allowed only four tries per level, meaning they could not cycle back to their original version. Because we restricted their attempts, grading was also adjusted; instead of getting full points for 100% performance and requiring mastery, each student received the best score of their four attempts. By doing this, students were deliberately making a choice to try again for a better score but were not penalized if they tried again and performed less well. Thus, the threshold for moving onto the next level was relaxed to be 100% performance *or* attempting all four versions. This change was intended to give students more agency over their learning to improve ownership of the material, since subsequent versions after their first attempt would not be forced upon them, but trying again would be a decision students made to get more points or better learn the material.

Because of this more flexible mechanism for moving on, those who could not master a level were still able to see later levels and receive some points without resorting to unproductive behavior. Additionally, single weekly homework assignments were split into two smaller biweekly assignments to reduce student fatigue and encourage time management. Many students in the previous year complained that they could not finish the homework because they started too late, which probably contributed to their frustration and desire to complete the exercises quickly rather than use them as intended. We hoped that these changes would mitigate the desperation students felt which may have informed students' unproductive behaviors.

To confront student attitudes directly, we included an activity in the first discussion class that asked students to consider and record pros and cons of different types of homework styles, including mastery. Students worked individually and then shared ideas at their tables, then shared in a class-wide discussion. By allowing students to come up with reasons that the mastery homework could be useful to them and voice potential frustrations, the activity aimed to impress the function of homework as a learning tool, created to help them.

The homework's content itself was modified to address mismatched levels with the intention to create more uniform versions across each specific level. By looking at student performance data on the question, version, and unit level, we identified and adjusted problems that were unintentionally too difficult or too easy. Moreover, new levels were added or old levels were split to target student weaknesses revealed by the data with more scaffolding. This served the double purpose of helping students with difficult topics and ensuring that no students could pass through the homework without confronting their weaknesses.

In particular, one math competency level was added to most weeks. The math levels were designed to have students first practice the skills in a math context, then mirror the same process in a physics context, where the conceptual physics work was already completed and just the final math execution was left. Because we suspected that students' difficulty was in recognizing when to apply math skills to problems (Activating the tool, in the ACER model [126]), modeling the use of a familiar skill in a math context and repeating in the physics context was done deliberately to help cue students to required tools. An example of one version of a level on systems of equations is shown in Figure 6.1. This was the math level added to support two-body Newton's Second Law problems, the level described in the previous chapter as one which students were able to set up equations but not solve.

Math Supplement: Systems of Equations

Level 1 Competency:
i) To be able to solve a system of equations using substitution or elimination.

Given the system of equations,

$$\begin{cases} x + 3y = 23 \\ -3x - 2y = -6 \end{cases}$$

Solve for x and y.

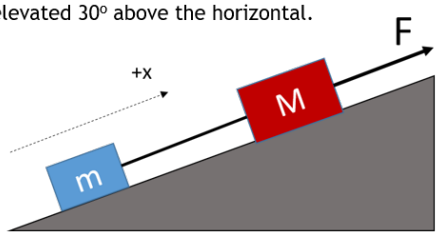
1) x =

Submit

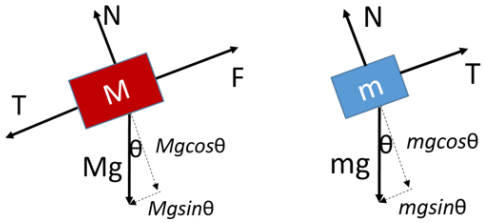
2) y =

Submit

3) Two blocks are connected by a string, and are being pulled up a ramp by a force F = 75 N, as shown. Their masses are m = 3 kg and M = 6 kg, respectively, and the ramp is elevated 30° above the horizontal.



To solve for acceleration, we need to draw free body diagrams for each block and write each block's equation of motion from Newton's 2nd Law. The free body diagrams look like this:



The blocks' acceleration is going to be in the x-direction (defined to be up the ramp, in the original picture), so we'll use the forces in the x-direction to write down the system of equations,

$$\begin{cases} F - T - Mg\sin\theta = Ma \\ T - mg\sin\theta = ma \end{cases}$$

Using this system of equations, solve for acceleration and the magnitude of the Tension.

The acceleration, a =

m/s^2 Submit

4) The magnitude of the tension force, T =

N Submit

Figure 6.1: Example version of a math level on systems of equations, designed to support students in solving two-body Newton's 2nd Law problems. The level begins with the math skill in a familiar math context, repeated in a physics context.

Over seven weeks, 30 content levels from Fall 2014 became 36 levels in Fall 2015, and 522 questions were expanded into 684 questions. A list of levels for both 2014 and 2015 is listed in Table 6.1. Two weeks (week 3 and week 7) were left unchanged to enable comparison from the first year to the second; the levels in Week 3 were generally achievable, but Week 7 was very difficult for students. Additionally, feedback was added for the most commonly chosen wrong answers, so that instead of only showing the correct way to work a problem, we could also target students' specific conceptual mistakes.

Table 6.1: List of levels from 2014 and 2015. Levels were given as one assignment in 2014 and broken into two biweekly assignments for 2015.

Topic	Week	2014	2015
Kinematic Definitions	Week 1a	L1: v-t graphs L2: x-t graphs L3: Multiple-body graphs	L0: Warmup L1: v-t graphs L2: x-t graphs
	Week 1b	*N/A*	L3: Math supplement: Coordinate geometry L4: Multiple-body graphs
Constant acceleration and 1D relative motion	Week 2a	L1: Relative motion in 1D L2: 1D Constant acceleration 1 L3: 1D Constant acceleration 2 L4: 1D Constant acceleration 3 L5: 1D Constant acceleration 4	L1: 1D relative motion L2: Constant acceleration L3: Free fall
	Week 2b	*N/A*	L4: Math supplement: Coupled equations L5: Constant acceleration with symbols
Vectors and 2D relative motion	Week 3a	L1: Vectors L2: 2D Relative motion 1 L3: 2D Relative motion 2	L1: Math supplement: Vectors
	Week 3b	*N/A*	L2: 2D Relative motion 1 L3: 2D Relative motion 2
Projectile motion	Week 4a	L1: Projectile motion 1 L2: Projectile motion 2 L3: Projectile motion 3	L1: Projectile motion 1 L2: Projectile motion 2
	Week 4b	*N/A*	L3: Math supplement: Algebraic relationships L4: Projectile motion 3
Newton's second law	Week 5a	L1: Drawing FBDs L2: Equations from FBDs (Ramps) L3: Equations from FBDs L4: Full problem L5: Newton's second law with two objects	L1: Drawing FBDs L2: Equations from FBDs L3: Math supplement: Rotated coordinate systems
	Week 5b	*N/A*	L4: Weight vector in a rotated˜coordinate system L5: Equations from FBDs (Ramps) L6: Full problem
Newton's first and third laws	Week 6a	L1: Identifying Newton's third law pairs L2: Newton's third law (static) L3: Newton's third law (accelerating) L4: Introduction to friction (as a constraint force)	L1: Math supplement: Systems of equations L2: Newton's second law with two objects L3: Identifying Newton's third law pairs
	Week 6b	*N/A*	L4: Newton's third law (static) L5: Newton's third law (accelerating) L6: Introduction to friction (as a constraint force)
Friction and uniform circular motion	Week 7a	L1: Static friction L2: Kinetic friction L3: Static or kinetic friction? L4: Uniform circular motion	L1: Static friction L2: Kinetic friction
	Week 7b	*N/A*	L3: Static or kinetic friction? L4: Uniform circular motion
Universal gravitation and springs	Week 8a	L1: Gravitational forces L2: Orbital motion L3: Springs	L1: Math supplement: Ratios L2: Gravitational forces
	Week 8b	*N/A*	L3: Orbital motion L4: Springs

6.3 Results

6.3.1 Student Frustration

The main objectives for the implementation of mastery-style homework in its second year were to ease frustration and improve student engagement with the system. The effect of the changes to the content and delivery dramatically reduced the students' tendency to become more frustrated as the term went on, as well as improved students' original perception of mastery-style homework. As in the previous year, students self-reported their response to the mastery homework in an online survey, rating their interactions with the system on a Likert scale from "Very Encouraging" to "Very Frustrating;" a normalized histogram of the results of the survey question for 2015 are shown in Figure 6.2, also with results from 2014 for comparison.

Though students still found the system more frustrating as time went on, the shift was much less dramatic. By week 5, the portion of students who reported feeling frustrated was only about 30%, halved from 60% in 2014, and comparable to students' initial frustration in 2014. Additionally, students began the semester feeling less frustrated than their classmates did in 2014. The improvement in students' initial impression of the homework could be attributed to the priming exercise in the first class. Using the average reported scores, the effect size in reducing frustration from the first year to the second year was 0.8 with $p < 0.001$.

Student responses to the open-ended prompt attached to the survey showed that many students saw the mastery homework as something that was useful for them, and cited mastery as their preferred type of homework:

> "I really liked it because it kind of seemed like your own personal teacher that explained the materials to you that you did not understand. I really liked this version of homework especially for me coming into this class with no physics background its providing me with that extra help you usually don't find in college and our hectic schedules. I would prefer that homework was set up this way than the other versions."

> "Mastery homework is probably my preferred method of homework completion, since it supplies the user with multiple tries and provides video assistance to ensure, well, mastery, of the subject at hand."

Responses also indicated that students saw the value in repeating problems, saw their learning as something flexible and growing, and believed in their ability to improve through the system and gain confidence. Students also appreciated an ongoing assessment of their current understanding, which can sometimes be lost in unlimited try style homework:

> "I like the fact that it highlights the important parts and lets me see just how well I know the material, instead of how well I think that I know it."

> "The mastery homework is extremely helpful in encouraging me to understand the material. In the past, getting problems wrong on homework was very discouraging because I was not able to fix my mistakes in order to get a higher grade. With the mastery homework, I am able to fix my mistakes and get a better grade by trying my hardest to truly understand the concepts."

> "I like that I am not penalized for my mistakes and can learn the material immediately after."

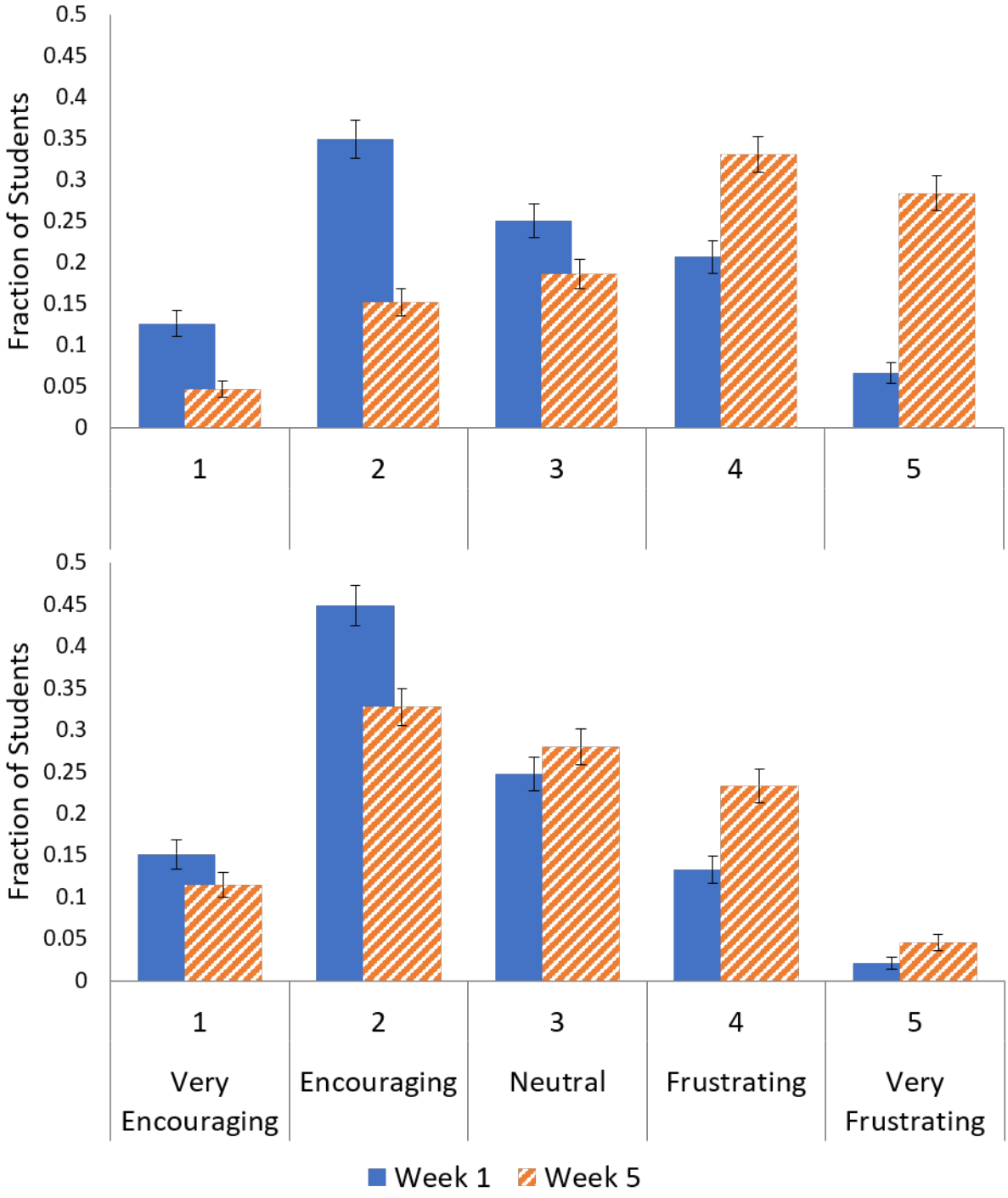


Figure 6.2: Normalized frustration histograms from (a) Fall 2014 and (b) Fall 2015: Students responded to the question, "Which statement best reflects your experience working through the mastery homework?"

> "I believe that the homework is very helpful for my learning. Having to not worry about being right about every problem releases the load of worry and adds more focus to my studying. Plus, we learn better by doing things the wrong way. We learn to see problems from different perspectives. Thank you for teaching this way. It is very helpful in my opinion."
>
> "I have found that mastery homework has given me the challenge of doing problems I do not immediately know the answer too, but it also has given the benefit of an immediate explanation followed by more problems to practice my new knowledge on. It is a great system and I love it."
>
> "The mastery homework has helped tremendously with practicing physics problems. The multiple tries helps, but I think what helps more is the fact that if I try a new problem cluster, it rearranges problems and gives me new numbers/set-ups so that I'm constantly trying different situations. This approach to physics causes me to always solve various types of problems, which is great for getting better."
>
> "I like that it does force you to be confident with your answers because there are only four tries total while giving you chances to make some errors, find out what you did wrong by watching the solutions, and correcting the problems to feel better about comprehension of the material."

Many students also expressed that the solution videos were useful to them, both in the Week 1 and Week 5 survey. As this was a problem in the previous iteration, the positive feedback was encouraging:

> "Whenever I got questions wrong, I was able to view the videos to learn from my mistakes and do it better the next time. It was easy to learn that way."
>
> "I have found video solutions to be more helpful than written out solutions."
>
> "Having detailed solutions to problems that I got incorrect is immensely helpful."
>
> "I loved the solutions provided for some of the problems. It really elucidated some of the concepts that I was getting wrong, or did not understand."

Students did still have frustrations with the system, particularly the amount of time spent and redoing all problems after missing one. The second complaint was one that was prominent in the previous iteration of the homework, and we again saw that frustration with this aspect increased as the semester continued. Some students still found the versions disconnected. However, many of the frustrated responses showed that students understood the repetition as part of the learning, even if it was frustrating:

> "The only negative I have about the mastery homework is that it takes a long time to do. However to master anything you must take the proper amount of time to understand the material you are trying to learn."
>
> "I would make the questions more similar. For example if you miss a question with west as the positive direction then give a different question with west still as the positive direction."
>
> "It is only frustrating to have to redo the entire 5 problems again even if we only missed one and it is also becomes frustrating to receive a new set of problems that appear to be much more difficult than the previous ones."

> "It gets annoying to do an entire problem set and make one small mistake then have to redo it all to get full credit but other than that it's a fine way of doing things."

The limitation of only four attempts, which was new for this year, was also a concern to some students, and they expressed interest in having more or unlimited versions:

> "I think having unlimited try is the best way to learn. Not because of grades, but because I can figure out what I can fix right away"

> "I think the solutions are really helpful, but I wish there were more tries because sometimes it takes a little practice before I can fully understand and get it right."

> "The mastery homework was helpful in that it provided quick enough feed back once you submitted all your answers and had solution videos. A downside is that I wish students had five attempts instead of four to receive a 100% being that I struggled with one type of problem many times."

> "The mastery homework is beneficial, however being restrained by the 4 attempts does not help when a concept is difficult because it takes multiply tries on a difficult problem to finally get it all right with all parts."

Some of the aspects of mastery homework which frustrated students in the first implementation were still present, but overall responses showed greater thought to the benefits of the system and negative responses were often couched in softer language. Students often cited the solution videos as being useful to them, and spoke of the homework more often as a means to learn, recognizing the process as a growth opportunity rather than an obstacle. Having students brainstorm on the role of homework and potential benefits of types of homework in their first class may have helped students articulate what they value in the system, which was communicated through this open feedback.

6.3.2 Student Engagement

One expects that from improved attitudes, improved behaviors will follow. Although "skipping" was technically not feasible anymore with non-repeating versions, a look at the time distributions on the levels shown in the previous chapter lets us confirm that the number of students spending under a minute on their fourth attempt (while not mastering) was reduced significantly. Histograms of time spent by students on their final attempt are shown in Figure 6.3, with the distributions from 2014 included for comparison. Note also that there were fewer students who did not master on their fourth attempt, suggesting that students worked more earnestly on each of their attempts and either mastered on a previous version or on their fourth attempt.

The data revealed that mismatch between versions of levels was still a problem. Using the same definition as before, a level is considered having mismatched versions if either the overall mastery rate between versions or at least two questions showed a difference of 20% performance. By this definition, 25 of the 36 levels were classified as mismatched, compared to 20 of 32 in the previous year, suggesting that our efforts were unsuccessful. There were fewer levels which showed differences in mastery rates, but questions still had a lot of variation. This could in part be due to the fact that levels with fewer than 4 versions in the previous

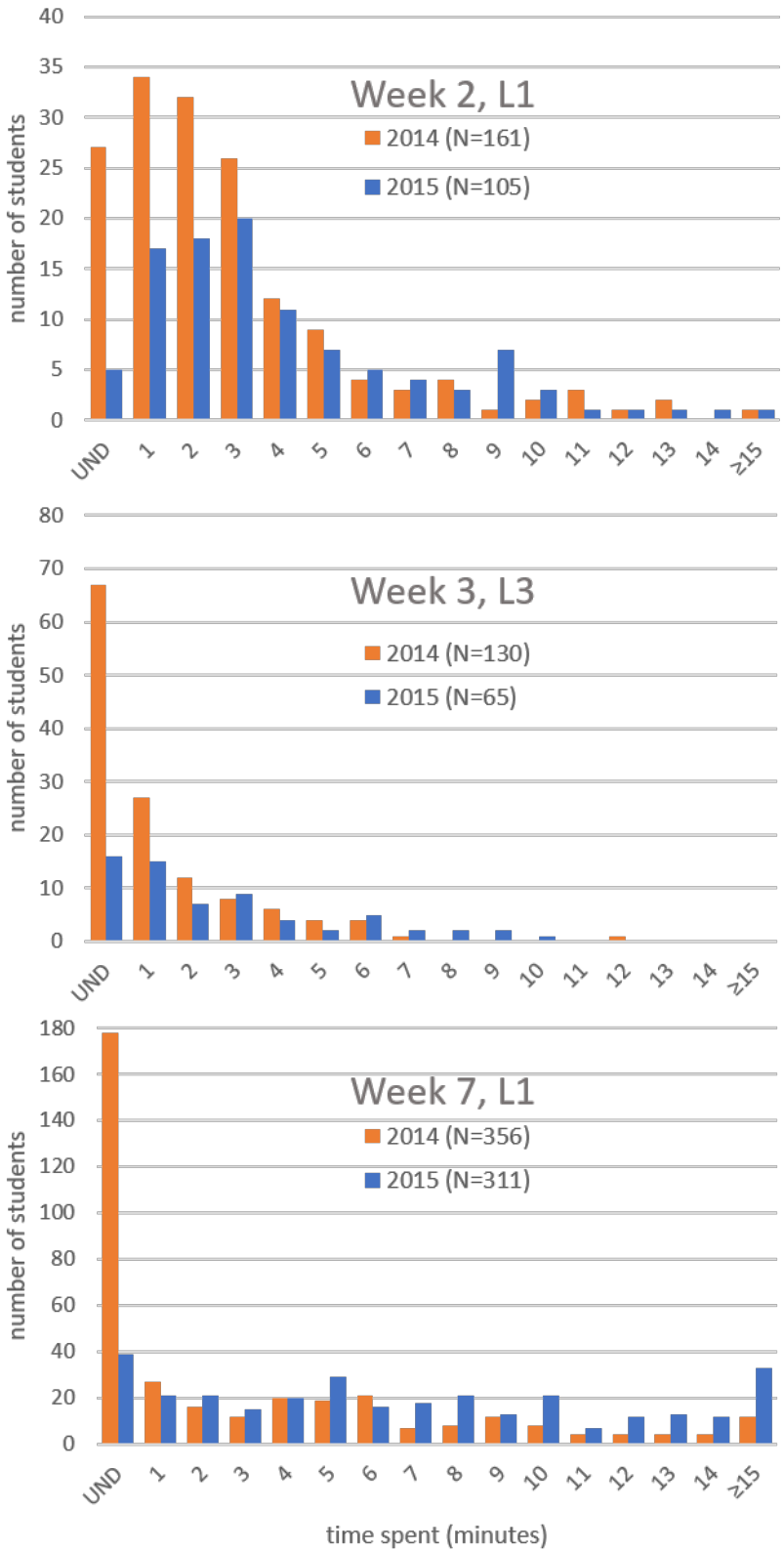


Figure 6.3: Histogram of time spent on students' fourth attempt, if they did not master on that attempt, for both 2014 and 2015.

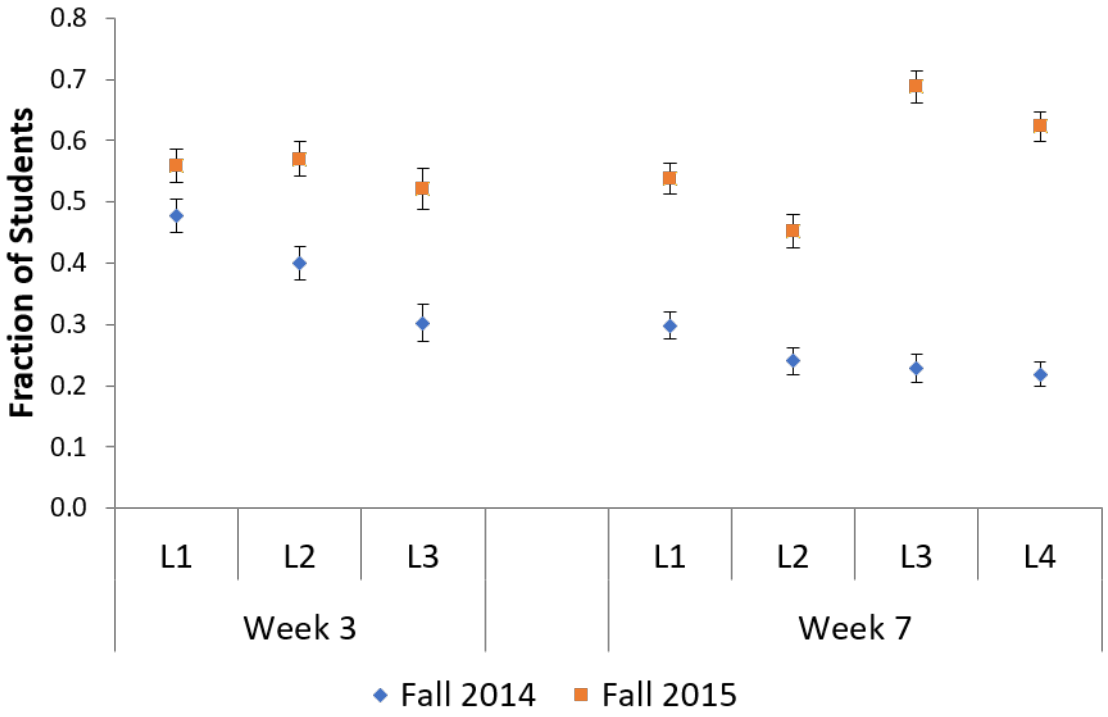


Figure 6.4: The fraction of students who watched over half of the solution videos to their missed problems, separated by levels in two unchanged weeks from 2014 to 2015.

year were expanded to have a full set of 4 versions; some of the mismatch increases were in levels which only had two versions in 2014 and so the "matching" condition was more easily achieved in the previous year. Additionally, many new levels were introduced which we had no prior data for.

Despite the persistence of mismatch in students' performance from version to version, time spent on solution videos significantly increased compared to the previous year. Figure 6.4 shows the fraction of students who watched more than 50% of the solutions to missed questions on the unchanged weeks. The threshold for viewing was 10% of the solutions' narration time. Although the content was not changed for weeks 3 and 7, students in Fall 2015 were more conscientious about using the solutions. The effect is most noticeable in week 7, where the average fraction of students choosing to watch solutions more than doubled.

6.3.3 Math Performance

From 2014 to 2015, there were 7 math supplement levels added or edited, and 4 of the 7 preceded levels which were consistent between the years to allow comparison. Table 6.2 shows a summary of all math levels given in 2015 and their targeted content.

Using these five supported levels' mastery rates (corresponding to four math supplement levels), one can see improvement in performance for students with the math supplements via their mastery rates in Figure 6.5. The level for multiple body graphs only had three distinct versions in 2014, so only tries 1-3 are shown for that level, and the rest are shown through four tries corresponding to four unique versions. Some

Table 6.2: A list of math supplement levels and level(s) they intended to support in 2015, with notes on differences in the supported levels between years. Levels which were similar in both years and had a new math level introduced are able to be compared for analysis of the math level's effect.

Math Level	**Following Level(s)**	**Notes on changes from 2014 to 2015**	**Comparison possible**
Coordinate geometry	Multibody x-t graphs	In 2015, the level gained one question from 2014.	✓
Coupled equations	Constant acceleration with symbols	The level of constant acceleration with symbols was not given in 2014. This level and its math supplement were added in response to student difficulty on exams.	
Vectors	2D Relative motion	There was a level on vectors in both 2014 and 2015.	
Algebraic relationships	Projectile motion 3	Projectile motion 3 was dramatically changed from 2014 to 2015.	
Rotated coordinate systems	**Newton's second law:** Equations from Ramp FBDs	The Newton's Second Law level changed from 2 questions in 2014 to 4 questions in 2015. New questions were isomorphic.	✓
Systems of equations	**Newton's second law with two objects**	This level was not changed from 2014 to 2015.	✓
Ratios	Gravitational forces; Orbital motion	These levels were not changed from 2014 to 2015.	✓✓

levels showed significant improvement from 2014 to 2015 with the addition of math supplement levels, while some showed little change. A summary of these mastery rates and their effect sizes with significance are given in Table 6.3. The effect sizes of levels which showed significant improvement ranged from 0.17 to 0.5, which are small to medium effects. The largest effects were in the two body Newton's Second Law level, which was a driving motivation for the addition of the math levels from 2014. In both the Equations from Ramp Free Body Diagrams and two-body Newton's Second Law levels, students with the math supplements outperformed their counterparts from 2014 on their first attempt, suggesting that students went into the problems stronger than students without the math levels.

These results may be inflated due to the changed delivery of mastery-style homework in 2015; mastery rates in 2014 could have been low due to students "skipping" until returning to a previously seen level. To address this, the analysis was also done removing students who "skipped," which we defined to be students spending under 60 seconds on a version, or 1/4 of the median time it took mastering students to complete the level. For all the levels except the Equations from Ramp Free Body Diagram level in 2014, the median time was above 4 minutes, so 60 seconds was used. In 2014, the median mastering time for the Equations from Ramp Free Body Diagrams was 79 seconds, so 20 seconds was used as the cutoff for skipping students. Note that this time is low because the level only had two questions in 2014. A more conservative comparison of the two groups' mastery rates, with skipping students removed, is shown in Figure 6.6. Using this subset of non-skipping students, the significance of the third try for Multiple body x-t graphs and Tries 3 and 4 of Equations from Ramp Free Body Diagrams reduced to no longer be significant, but all other results remained highly significant, suggesting that these effects are not solely due to the mastery delivery but also are a result of the new math levels.

Additionally, two of the studied levels gained questions in 2015, which made achieving mastery more difficult. The multiple body x-t graph level gained one question by requiring students to give equations of motion for both objects in 2015, while only asking for the more "difficult" equation of motion in 2014 (for example, the car or person who had a time delay or started at a position other than the origin). The level which asked students to write equations from Free Body Diagrams on ramps also increased from 2 to 4 questions in 2015; these additional questions were simply isomorphic versions of the sets of two questions. In 2014,

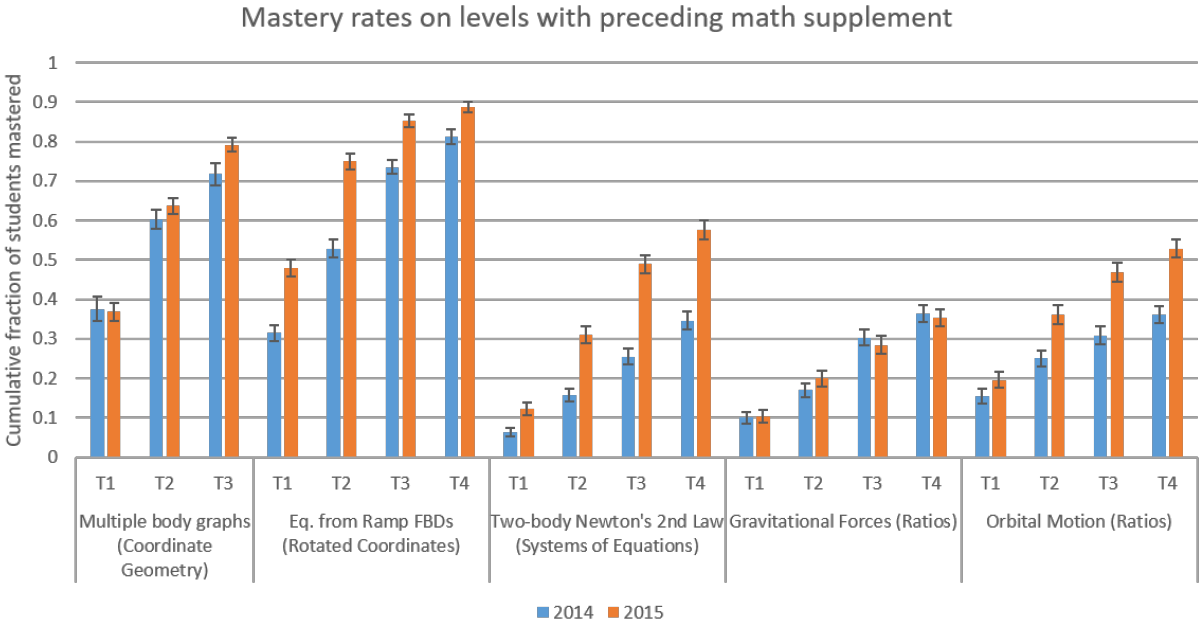


Figure 6.5: Cumulative mastery rates by Try (T1, T2, etc) for students in 2014 and 2015 on levels supported by math supplements in 2015. The math supplement level is given in parentheses.

Table 6.3: Summary of mastery rates by try in 2014 and 2015 on levels following math supplements, with effect sizes and p-value significance.

Level	**Try**	**fraction mastered**		**N**		**effect size**	**significance** (p value less than *0.01, **0.001, ***0.0001)
		2014	**2015**	**2014**	**2015**		
Multiple body graphs	1	0.376	0.369	500	495	–	
	2	0.604	0.637			–	
	3	0.718	0.792			0.17	*
Equations from Ramp FBDs	1	0.315	0.479	515	470	0.34	***
	2	0.529	0.749			0.47	***
	3	0.735	0.852			0.29	***
	4	0.812	0.888			0.21	**
Two-body Newton's Second Law	1	0.064	0.122	497	457	0.20	*
	2	0.158	0.311			0.37	***
	3	0.255	0.489			0.50	***
	4	0.347	0.577			0.48	***
Gravitational Forces	1	0.100	0.104	460	417	–	
	2	0.170	0.200			–	
	3	0.303	0.285			–	
	4	0.365	0.353			–	
Orbital Motion	1	0.154	0.196	427	410	–	
	2	0.251	0.361			0.24	**
	3	0.309	0.469			0.33	***
	4	0.361	0.529			0.34	***

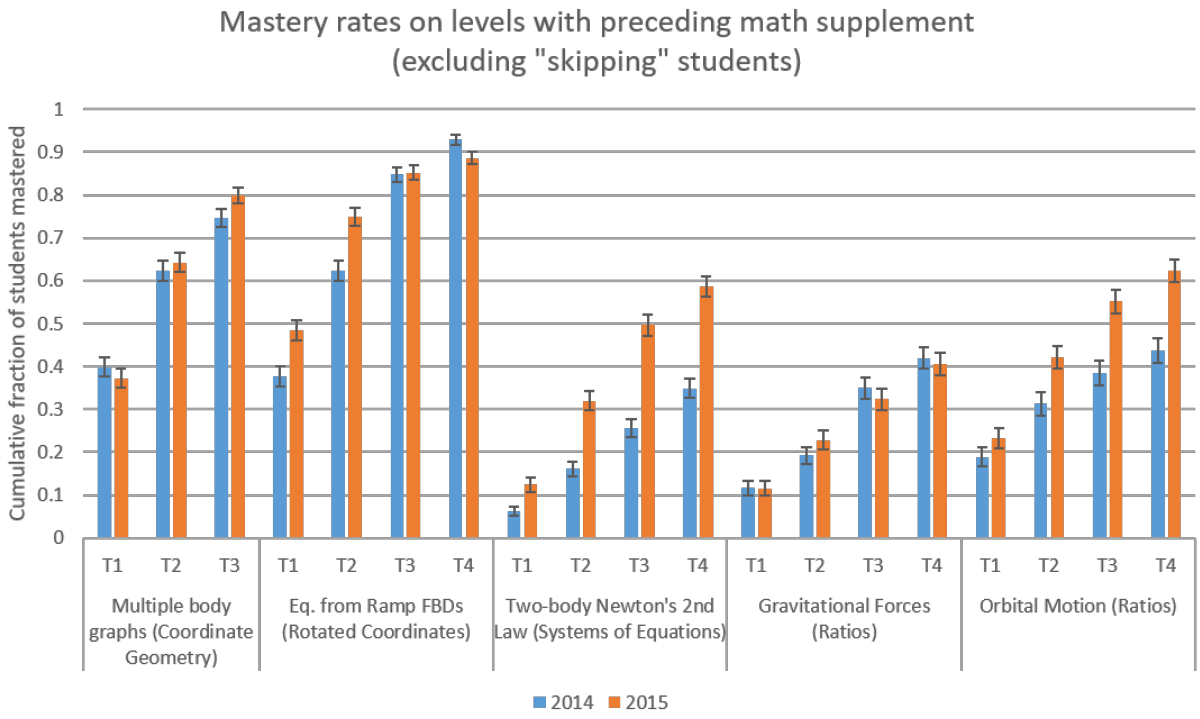


Figure 6.6: Cumulative mastery rates by Try (T1, T2, etc) for students in 2014 and 2015 on levels supported by math supplements in 2015, excluding students who spent under 60 seconds on their try (or under 20 seconds for students in 2014 on the Eq. from Ramp FBDs level). The math supplement level is given in parentheses.

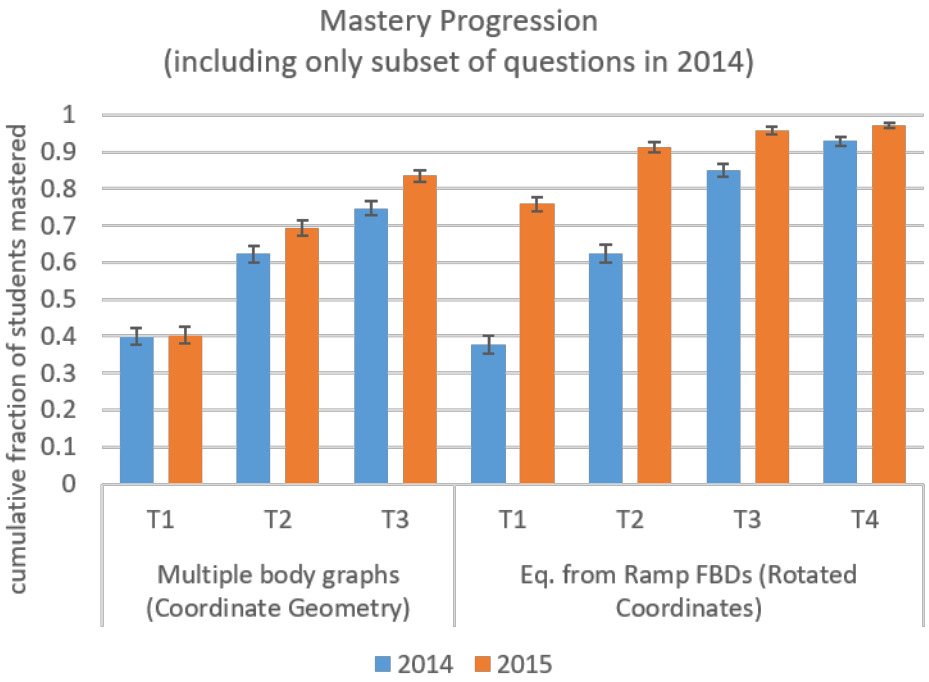


Figure 6.7: Fraction of students mastering in both 2014 and 2015 if mastery is defined by only subset of questions given in 2014. This plot only includes students who were not classified as "skipping."

students were given one free body diagram and asked to write the equations of motion in the (rotated) x and y directions. In 2015, students were asked to write two equations for two different ramp scenarios, essentially doubling the level. This was done to reduce the possibility for students to guess correctly without mastering the material. For a more fair comparison, Figure 6.7 shows mastery rate progression for students in 2014 and 2015, with 2015 mastery being defined as correctly answering the corresponding questions that were given in 2014. By this definition of mastery, the third try of Multiple body graphs again became significant (with an increased significance of $p<0.001$). The effect sizes nearly all increased, with large effect sizes of 0.83 and 0.73 for the first and second attempts at the Newton's Second Law ramp problem. The only effect size which did not increase was the fourth try of Newton's Second Law, which had already reached a mastery rate of over 90% in its second try and saw a ceiling effect while students from 2014 approached high mastery rates. Table 6.4 shows the effect sizes and mastery rates for students on these changed levels if we assume answering the same subsets of questions constitutes mastery.

Recall that a motivation for the math levels was in response to a level on Newton's Second Law with two bodies where students could correctly set up the equations of motion for the problem but under half of the students could correctly solve. To check changes in students' ability to calculate after having the full set of equations, this level was analyzed again. Questions 1-4 asked students to (1-2) write equations of motion for each object and (3-4) specify that the relationship between the accelerations of each (equal in magnitude, sometimes opposite in sign) and the magnitude of tensions (equal magnitude). Question 5 asked students to solve the system of equations for the objects' shared acceleration.

Considering only the pool of students who correctly answered questions 1-4, the fraction of students correctly

Table 6.4: Fraction of students mastering in both 2014 and 2015 if mastery is defined by only subset of questions given in 2014, with effect sizes and significance. Note that the mastery rates and number of students are different from Table 6.3 because this analysis only includes students who were not classified as "skipping."

Level	**Try**	**fraction mastered**		**N**		**effect size**	**significance** (p value less than *0.01, **0.001, ***0.0001)
		2014	**2015**	**2014**	**2015**		
Multiple body graphs	1	0.399	0.403	451	472	–	
	2	0.623	0.694			–	
	3	0.747	0.835			0.21	**
Equations from Ramp FBDs	1	0.377	0.758	492	454	0.83	***
	2	0.623	0.913			0.73	***
	3	0.848	0.958			0.38	***
	4	0.929	0.972			0.20	**

calculating from their correct equations are shown for 2014 and 2015 in Figure 6.8. These values are not cumulative, but represent fractions of the number of people on each of their tries who satisfy the correctness condition for questions 1-4 (the try number corresponds to their try at the level, not their number of tries after getting the equations correct). Overall, the number of attempts with correct equations in 2014 were 299 corresponding to 250 students. In 2015, there were 432 attempts corresponding to 348 students. A larger fraction of all students in 2015 were able to correctly set up equations (0.74 versus 0.51), and a larger fraction of those students were able to complete the calculation. Overall, the fraction of correct calculations (of those setting up correctly) saw a gain of 15.7%, corresponding to an effect size of 0.32. Considering each student once (rather than weighting them by their number of tries), the fraction of students who could ever correctly solve after setting up the equations correctly, saw a gain of 21.6%, which corresponded to an effect size of 0.48. Both are medium effects and were statistically significant with a p-value of $p < 0.0001$.

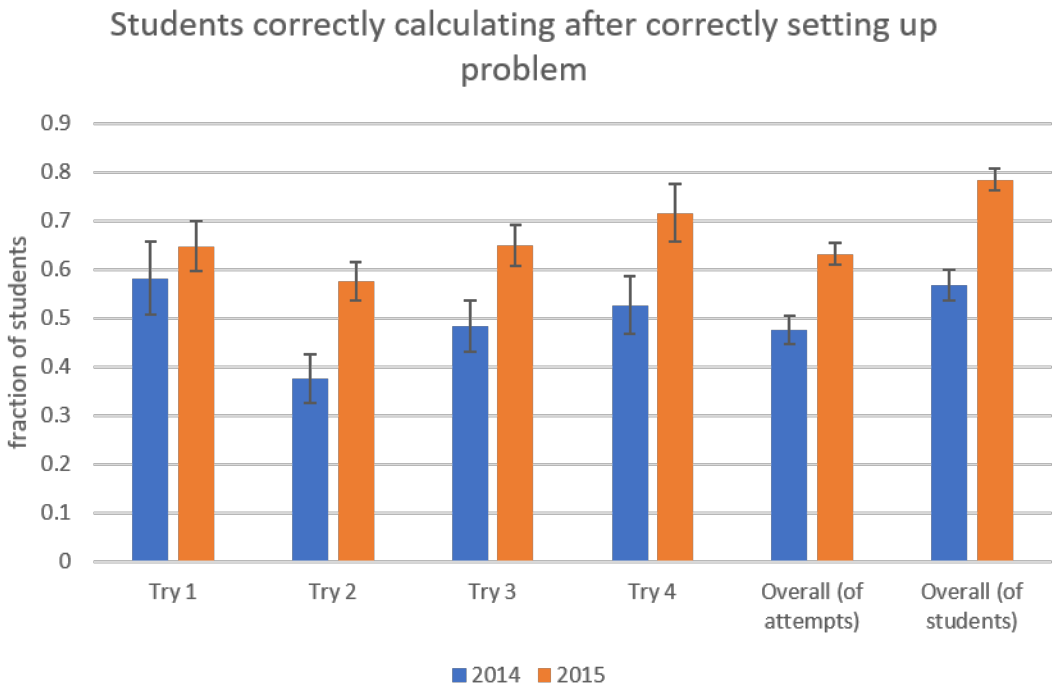


Figure 6.8: Fraction of students who correctly calculated after choosing correct equations of motion for both objects in the Two-Body Newton's Second Law level, by try. These are not cumulative fractions; students who attempted more than once are included in each try when they correctly set up the equations. The overall (of attempts) fractions are of all the attempts and contain multiple tries by individuals (about 17% of included students are included twice and 3% included three times). The overall (of students) fractions are the number of students ever correctly calculating of the students who correctly set up equations, and do not contain duplicate entries for students.

6.3.4 Student Performance

Looking at the number of attempts required for students to master levels (100% performance) more generally provides information of how students are interacting and learning from the activities. The culmination of the changes made affected student behavior, which is reflected in more serious attempts by students during the second year of implementation. Figure 6.9 shows the fraction of students who mastered each level by attempt for the two weeks with identical content in 2014 and 2015. Note that the mastery rate on the first attempt is consistent between the two years, suggesting that the groups of students are comparable. For example, the average difference between first attempt scores from 2014 to 2015 in week 3 was 1.7% $\pm$ 1.8%, and 1.8% $\pm$ 1.1% in week 7. The differences in the groups' baseline understanding were not significant, but students mastered levels much more rapidly in 2015, for easier and more difficult levels. These levels had mastery rates ranging from below 30% to above 80%, and all saw improvement.

Evaluating student performance over the entire semester, the average fraction of levels mastered by students (in the first four tries, before repeated versions in 2014) was systematically higher in 2015 than in 2014, shown as a function of week in Fig. 6.10. The scatter plot shows the weeks' average fraction of mastering students with Fall 2014 on the horizontal axis and Fall 2015 on the vertical axis; every point is significantly above the line, which indicates improvement of students in 2015 over students in 2014. Integrating over the all weeks, this effect is medium and statistically significant with an effect size of 0.5 and p <0.0001.[2] Again, improvements in mastery rates were seen for all levels regardless of difficulty.

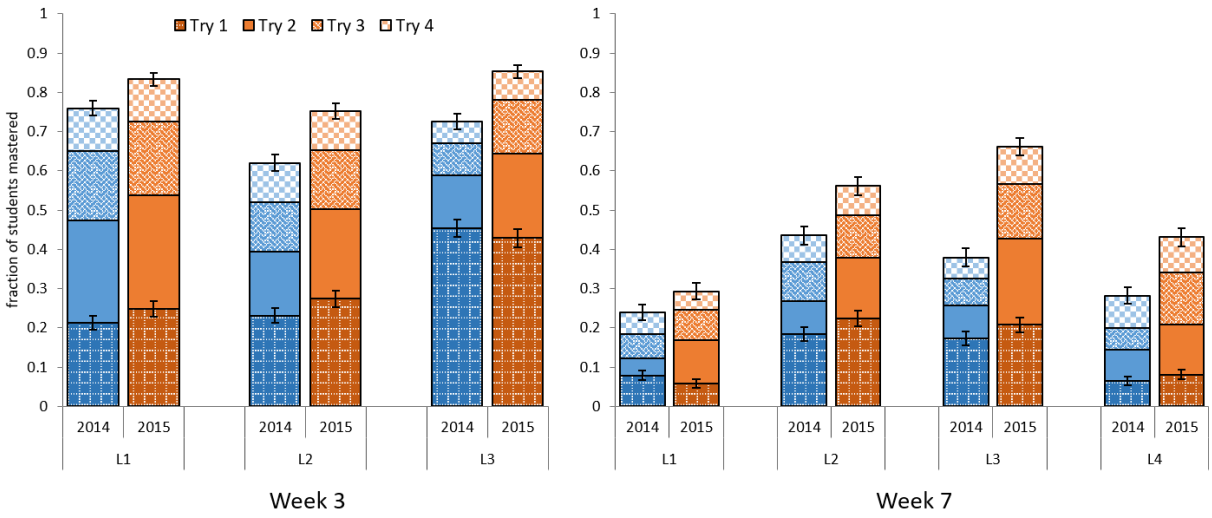


Figure 6.9: Mastery progression for students by try in unchanged weeks (Weeks 3 and 7) for 2014 and 2015.

[2]This effect size was cited as 0.7 in the publication "Mastery-style homework exercises in introductory physics courses: Implementation matters" in 2018 [11], which excluded Week 8 from the analysis. The exclusion of Week 8 was for simplicity because of the changing population of students after higher performing students dropped after the midterm. The low rate of mastery in Week 8 increased the standard deviation in this analysis, which led to a smaller effect size.

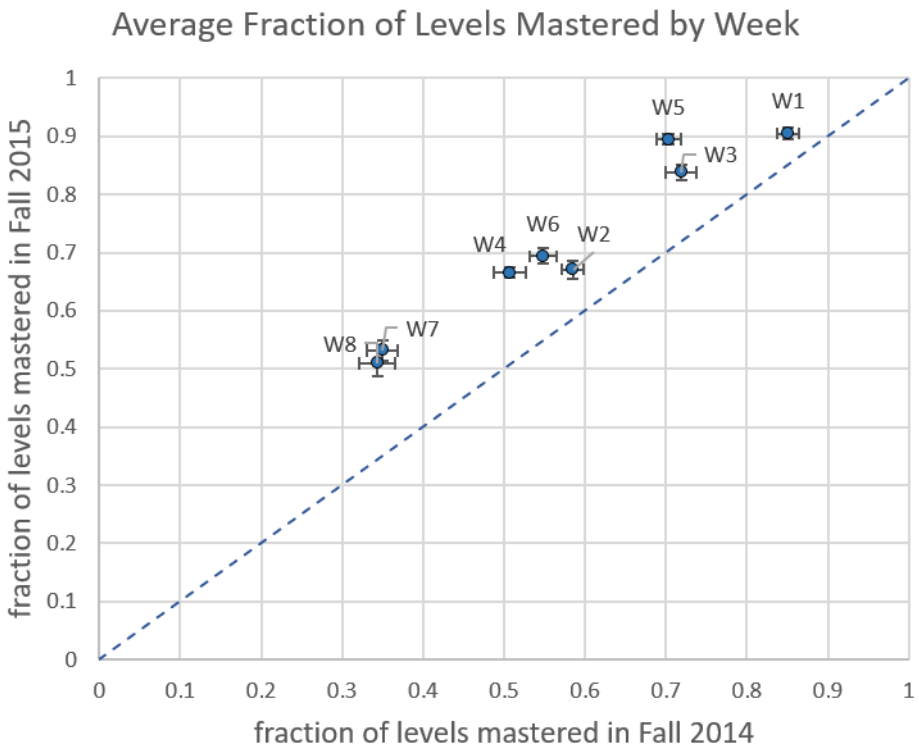


Figure 6.10: Average student mastery rates through semester by week: Both axes are the average fraction of levels mastered in each week, in 2014 and 2015. Points above the line indicate improved mastery rates for 2015 over 2014. This plot includes only students who attempted all levels, or all levels except one.

6.4 Discussion

The deliberate changes to the online mastery-style homework increased students' useful interaction with the system from its first year to its second year. On a small scale, disparities between versions were attended to by a combination of content reorganization and the adjusting of unintentionally problematic questions. These changes were not sufficient to reduce differences between versions, but there was more consistency in mastery rates between versions, even though some questions were still misaligned in difficulty. These changes, as well as the format change to a finite number of attempts, resulted in students' tendency to engage with solution videos more often. This suggests that students may have perceived an increased homogeneity of the difficulty of versions which made the solutions more useful, but it is more likely that the delivery method made each attempt more valuable. The direct affect conversation improved students' initial attitudes towards mastery and the attitudes continued to be more positive, signaling an increase in ownership and appreciation of learning which also appeared in students' open-ended feedback. Students were able to receive partial credit, but worked more sincerely for points even without the threat of receiving none for an unmastered unit, which is seen via improved mastery rates. The addition of math levels supported students' ability to succeed on targeted levels.

Unfortunately, it is not possible to completely isolate the impact of the individual changes. Unlike a clinical study where one has the luxury of prioritizing the research aspect, in this case we needed to prioritize our students. This required us to change all aspects that were likely to improve their learning. However, the data did allow us to infer that each of the components had some impact. The reduction in baseline frustration from one year to the next suggests that our direct affect conversation improved initial student perception of mastery-style homework. A comparison of weeks 3 and 7 (which did not change from 2014 to 2015) showed that students watched more solutions and mastered more often even when content was unchanged. By removing skipping students from our math level analysis, we saw improvement that cannot be attributed to the changes in the delivery method alone by providing a conservative comparison of only students engaged in both years. This suggests that students were taking the homework more seriously, rather than the material being simply more manageable. Still, this change in student attitude, behavior, and performance could be a consequence of the delivery style, our affect exercise, or a culmination of content changes in previous weeks. Technically, limiting students to four attempts deviates from pure mastery. However, by reducing the system's strictness, students felt a more compelling sense of agency and performed more as mastery theory prescribes, demonstrated by students' higher rate of mastering levels and using correctives (in our case, video solutions). Despite the relaxation of the mastery requirement, the changes amplified students' positive behaviors that take advantage of mastery learning principles. If one wanted to more clearly break apart the contributions to students' improved attitudes and behaviors, more clinical trials could be useful to untangle the effects.

When implementing mastery-style homework, the audience and its affective choices are crucial to consider. What was very successful for mature students in our clinical trial was much less successful for novices in a semester-long course. By targeting specific weaknesses, incentivizing positive behavior, and acknowledging student limitations, the system was more appealing to its users and thus, students engaged with it in a more productive way.

Chapter 7

Limits of Mastery: Algebraic Kinematics Scaffolding

7.1 Introduction

7.1.1 Limits of Mastery Overview

Moving forward with mastery-style homework in Physics 100, we were encouraged by the improvements from the first year to the second, but there was still space for evaluating the appropriateness of mastery in different contexts and addressing levels which had low mastery rates. Ideally, the homework content should be achievable with iterations over multiple versions, but 14 of 35 levels still had mastery rates below 60% after four tries. The fraction of students able to master each level within the four tries are shown in Figure 7.1, with levels of mastery below 60% in red.

Many of these levels were computationally complex or required students to perform multiple steps consecutively to correctly solve. Students' low mastery rates on these levels prompted further thought about which content may be appropriate for mastery and ways to adjust mastery for particularly difficult levels. Changes to the mastery homework are documented in the next two chapters; we added extra scaffolding to the problem prompt and problems[1] and looked at different grain-sizes of levels by breaking up levels with multiple steps into multiple levels. We also reverted some difficult levels into traditional immediate feedback homework, though the change did not have significant results and is not documented in this thesis. These changes were done over years 2016 through 2018. In each instance, the changes were given to half the course by splitting students randomly between two online courses that were otherwise identical.

[1]The work in this chapter on the effect of scaffolding was published in the Physics Education Research Conference Proceedings in 2017 as "Mastery learning in zone of proximal development" [136].

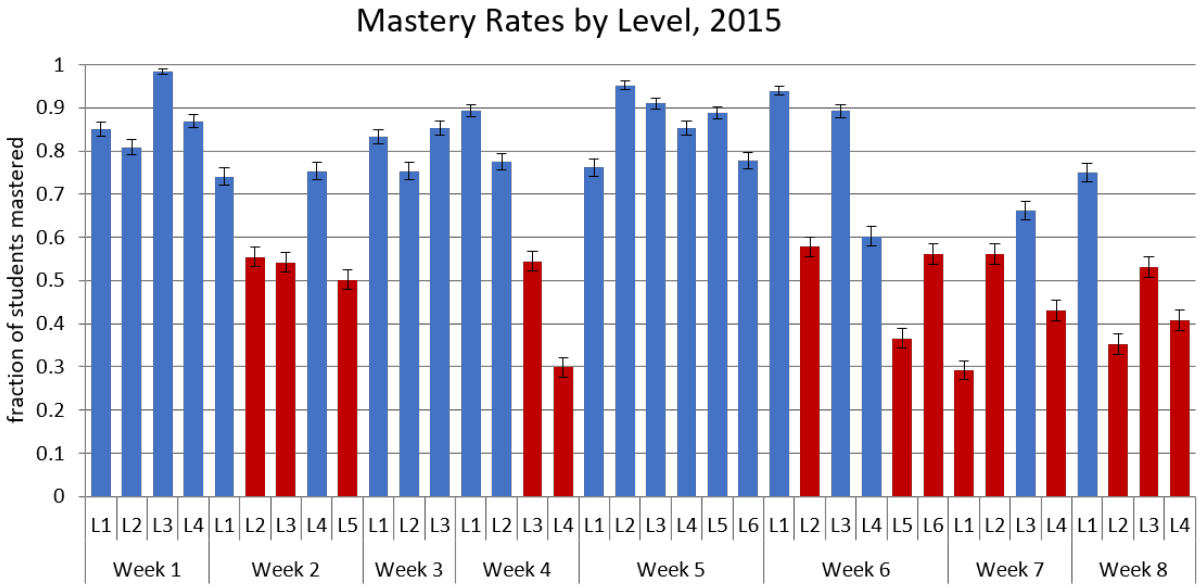


Figure 7.1: Fraction of students who were able to master each level within their four attempts. Levels with a mastery rate below 60% are indicated in red.

7.1.2 Algebraic Kinematics and Scaffolding

The use of math in physics problem solving is both an integral tool as well as a notorious barrier for some physics students. In particular, student difficulty with algebraic kinematic problems has been documented by Torigoe, Gladding and others, who published two studies in 2007 showing vast performance differences (up to 50%) on identical problems when numbers were replaced with variables [137, 138]. The expanded follow-up study showed that the performance difference was most dramatic for low-performing students [138].

The fraction of students who were able to master each level vary across the course's assignments, and a targeted training exercise for algebraic kinematics was very difficult for students to master in 2015 (Week 2, Level 5). This level had been previously added to directly confront student difficulty with algebraic kinematics, shown by consistently low scores on the type of problem on their midterm. Because this level is particularly difficult, it was a natural place to check for limits of mastery-style learning. While mastery learning has been shown to be appropriate in many instances [38], our course content occasionally challenges students beyond what they are able to master. Vygotsky's concept of the zone of proximal development (ZPD) encourages learning to take place in the space between what students can do and what they cannot do, within tasks that students can do only with assistance [32]. Low mastery rates imply that content is not in the range of skills most students can do, so introducing help may move student learning into the ZPD.

Scaffolding is a form of support directly intended to help students bridge the space between their current abilities and goals, which is often achieved by modeling or guiding students' practice [139, 140]. Scaffolding is rooted in Vygotsky's ZPD; effective support given by instructors should be in service of skills that are just beyond students' reach and should gradually be removed as student abilities improve [139]. A common type of scaffolding, classified by Beed et al as *strategic scaffolding*, involves guiding students through steps

to achieve an overarching goal [139]. By structuring and modeling the setup for students, they can focus on mastering the skills needed to complete the task without the cognitive load of creating the strategy. Scaffolding can also anticipate and address student errors or frustration, a benefit beyond the cognitive advantages [141, 142].

In 2016, we split the algebraic kinematics level into two variations with different amount of scaffolding, one with more guidance, intended to be easier to master, and one that left more work to the students themselves. The version with more guidance helped students identify the concepts and equations that were being used in each step, with a map of the overall steps to solve the problem given in the prompt. Due to some technical issues, we had a very limited sample size in 2016, so the experiment was repeated in 2017, which is documented in this chapter.

7.2 Methods

Two treatments were given separately by splitting all students in Physics 100 randomly between "two courses" online, both of which were identical up to Week 2's assignment. After four identical levels in Week 2, the fifth level differed for the two groups by providing "more" or "less" scaffolding for the same problems, followed by one more, new level (common to both groups), all delivered via the mastery system. The following level, Level 6, was not addressing the same competencies as the treatment in Level 5, but used the skills practiced in the treatment as a foundation to go further, so it was used as a proxy assessment to test how well the students could use the skills, albeit on a harder set of problems. The homework assignment also had a delayed feedback problem after the mastery levels. Students randomly received one of three versions assessing varying transfer from Level 6. The delayed feedback restricted students to one final answer for each question with feedback given only after the deadline, making the problem similar to an online quiz. The structure of these levels is shown in Figure 7.2.

In creating the versions of the algebra treatment level, the problems were intentionally meant to mimic the student processes required in Torigoe et al's studies and our difficult midterm problem. All versions had two accelerating objects and required students to plug one unevaluated algebraic expression into another algebraic expression. This was often solving for time in terms of variables and evaluating other characteristics of the two objects at that time to create a simplified ratio. Students who received "more" scaffolding had a strategy map included in the question prompt, which included the specific equations needed and correct signs on the variables. An example of one version of the type of problem is shown briefly in Figure 7.3 and detailed with different amounts of scaffolding in full in Appendix D.

Additionally, the extra scaffolding encouraged students to work in terms of a single object's variables; each step had multiple choice answers only in terms of object 2, for example. The "less" scaffolding versions gave the general kinematic equations in the prompt and provided small hints along the way, but did not provide specific equations or help students determine signs in the equations. The less scaffolded versions only reminded students to put everything in terms of a single object's variables at the end, which was necessary to simplify their ratio.

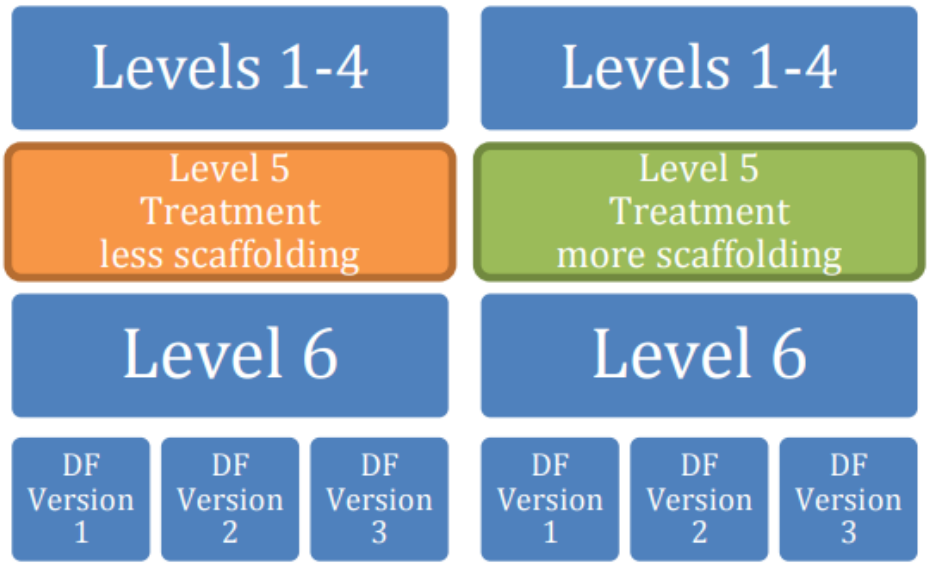


Figure 7.2: Schematic of treatment and assessment activities in Week 2 homework. Blue levels are identical for both courses, and students received different treatments (more or less scaffolding) for Level 5. Students were randomly assigned to one of the three delayed feedback (DF) versions, which vary in the amount of transfer from Level 6, with increasing transfer version 1, which was very similar to Level 6, through Version 3, which had the most transfer.

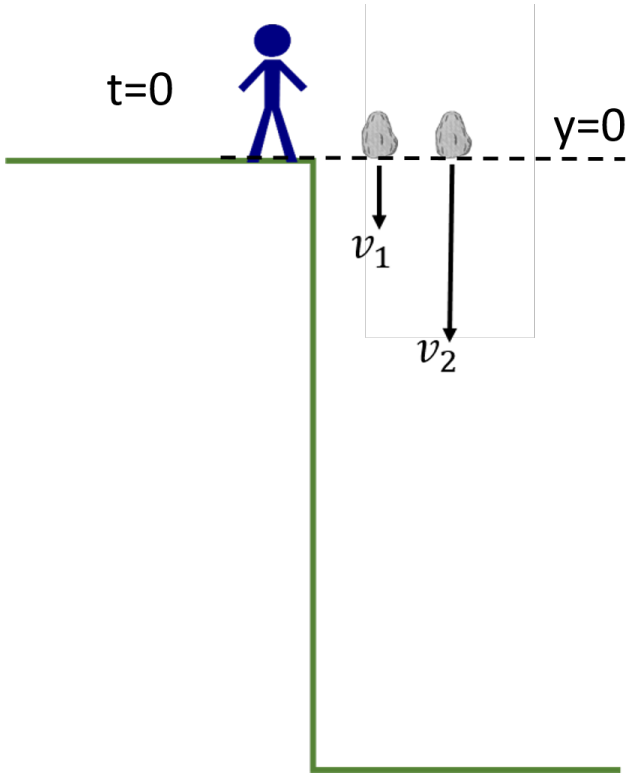


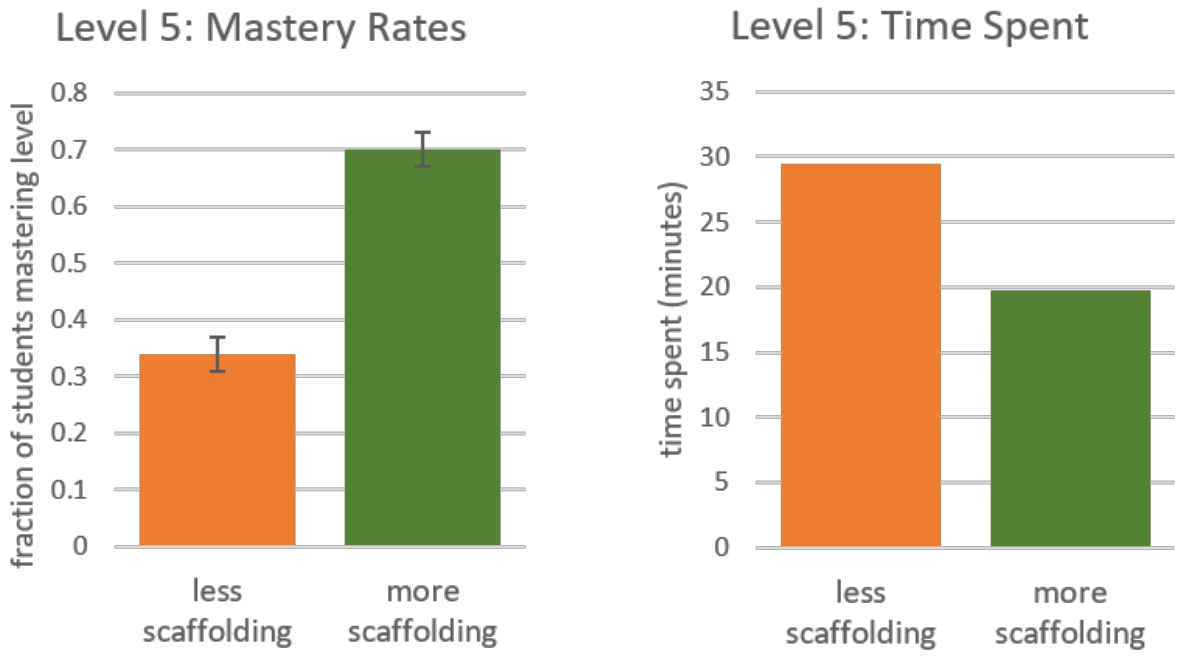
Figure 7.3: This version of Week 2 Level 5 serves the end goal of finding the ratio of the rocks' velocities when the faster thrown rock has fallen twice the distance from the top of the cliff compared to the distance fallen by the slower rock. The full problem with different amounts of scaffolding are shown in Appendix D.

7.3 Results

Just as in the previous year, levels within the mastery-style homework varied in difficulty and student mastery rates across the semester, but the different amounts of scaffolding in the split treatment affected students' performance and time spent. The additional scaffolding in the split treatment level successfully boosted the rate of mastery compared to students without extra training. The mastery rate doubled from 34% to 70% and the time spent on the level by those with extra scaffolding was significantly less, about 20 minutes compared to 30. Mastery rates and time spent for students on level 5 (where their treatments had different amounts of scaffolding) are shown in Figure 7.4.

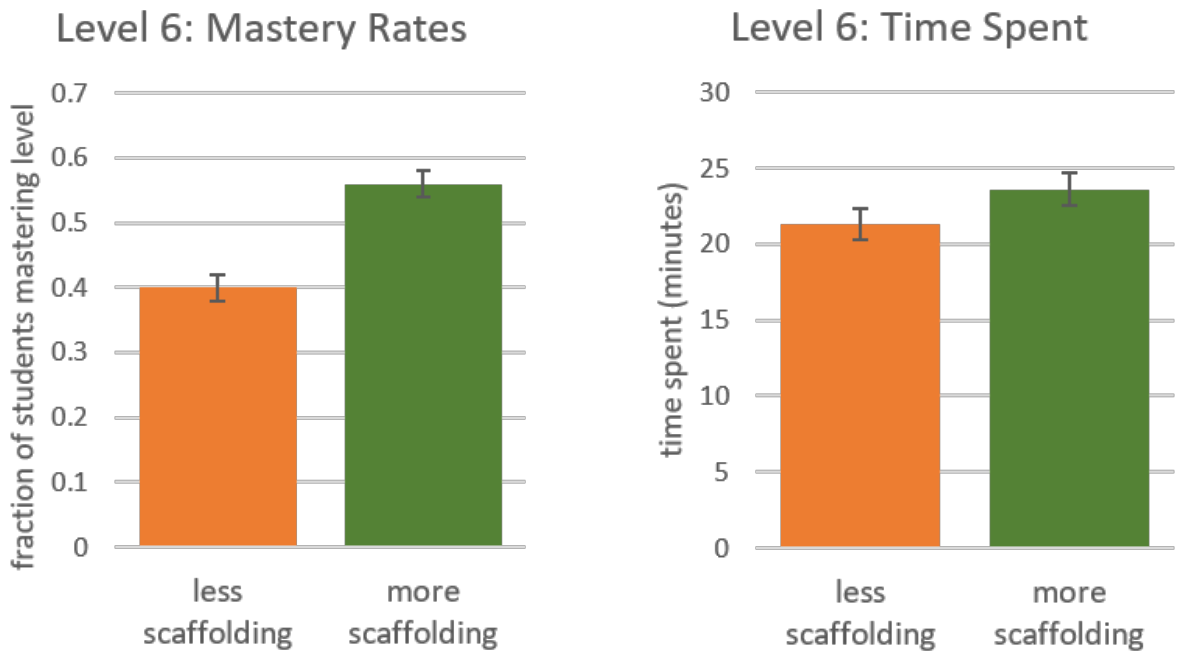
Performance measured via mastery rate on Level 6 (identical for both groups but more challenging than Level 5) was also higher for the students who previously had extra scaffolding, but both groups spent comparable amounts of time on Level 6. Mastery rates and time spent for Level 6 are shown in Figure 7.5. Recall that in the updated delivery system from 2015, students' final scores were their best score of their four attempts; students who saw extra scaffolding in Level 5 also scored higher on Level 6, meaning that they were answering more questions correctly on average. Mastery rates, score, and time spent on levels by both groups are shown in Table 7.1, with corresponding effect sizes and significance. Students who received more scaffolding were more likely to master (16% higher mastery rate, with $p < 0.0001$ and effect size of 0.65) and had a higher final score (8.5% score increase, with $p < 0.001$ and effect size of 0.36) on the same follow-up level.

Particular questions on the assessment were differently aligned with the treatment, so Figure 7.6 shows



(a) (b)

Figure 7.4: (a) Mastery rates and (b) time spent by students on Level 5. Level 5 had different amounts of scaffolding for the two groups of students.



(a) (b)

Figure 7.5: (a) Mastery rates and (b) time spent by students on Level 6. Level 6 was identical for both groups of students.

Table 7.1: Mastery rates, score, and time spent on levels 5 (different treatments) and 6 (identical for both groups). Differences are the values of "Less Scaffolding" subtracted from "More Scaffolding."

		Less Scaffolding	**More Scaffolding**	**difference**	**effect size**	**significance** (p value less than **0.001, ***0.0001)
Level 5	**Mastery Rate**	$34 \pm 3\%$	$70 \pm 3\%$	$36 \pm 5\%$	0.75	***
	Time Spent (min)	29.4 ± 1.2	19.8 ± 0.9	-9.6 ± 1.5	0.64	***
Level 6	**Mastery Rate**	$40 \pm 2\%$	$56 \pm 2\%$	$16 \pm 2\%$	0.65	***
	Final Score	$75 \pm 2\%$	$84 \pm 2\%$	$9 \pm 3\%$	0.36	**
	Time Spent (min)	21.3 ± 1.0	23.6 ± 1.1	2.3 ± 1.5	-	

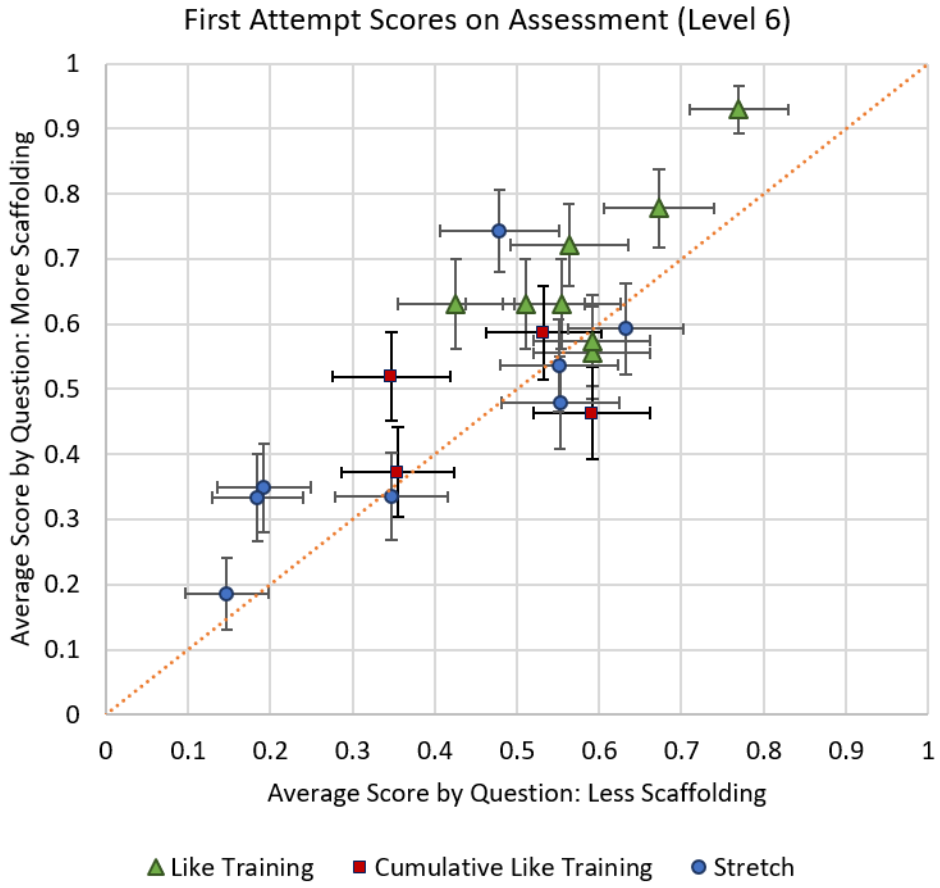


Figure 7.6: First attempt scores by question on Level 6, categorized by amount of transfer from Level 5.

question scores on students' first attempt, as well as categorization based on similarity to the skills in the treatment level. Questions that were "like training" corresponded to writing equations, "cumulative like training" was pulling the equations together to solve the system of algebraic equations, and "stretch" were conceptual questions that were not directly addressed in the treatment. Because Level 6 had five questions and four versions, there are 20 points on the scatter plot. Note that points above the line indicate better performance on the questions by students who had seen more scaffolding in their previous Level 5 training.

Using this grouping scheme, Table 7.2 shows the quantitative summary of the effects of the different treatments which showed up in each type of question in Level 6, cited as each type's average score. The scores of the delayed feedback levels following Level 6 are also given in Table 7.2. The delayed feedback questions are answerable up to a week late for 80% credit, so to remove those who were not taking their first attempt seriously, the values given are average scores on students' pre-deadline answer, after students spending fewer than 30 seconds were removed. Version 1 of the delayed feedback is most similar to Level 6, then versions are successively further transfer from the level.

Within Level 6, the largest difference was seen in "like" problems, with an effect size of 0.28, statistically

Table 7.2: Scores on different types of problems within Level 6, depending on transfer, and scores on the delayed feedback, with increasing transfer from Level 6, from Version 1 which was nearly identical to Level 6, to Version 3, which had the most transfer.

		Less Scaffolding	**More Scaffolding**	**difference**	**effect size**	**significance** (p value less than *0.01)
Level 6	**"Like" Questions**	$59 \pm 2\%$	$68 \pm 2\%$	$10 \pm 3\%$	0.28	*
	"Cumulative Like" Questions	$46 \pm 4\%$	$49 \pm 4\%$	$3 \pm 5\%$	-	
	"Stretch" Questions	$39 \pm 2\%$	$44 \pm 2\%$	$6 \pm 3\%$	-	
Delayed Feedback	**Version 1 (little transfer from L6)**	$48 \pm 3\%$	$58 \pm 3\%$	$10 \pm 4\%$	-	
	Version 2 (some transfer from L6)	$55 \pm 2\%$	$59 \pm 3\%$	$9 \pm 3\%$	-	
	Version 3 (most transfer from L6)	51 ± 3	51 ± 3	0 ± 4	-	

significant to $p < 0.005$. In the delayed feedback, the largest difference was in the most closely correlated version, with effects dropping off as the versions increase in transfer from Level 6. The strongest signal, for Version 1, had an effect size of 0.45 but a p-value of $p < 0.05$. Although splitting students into six sub-groups limited our data's significance, the results are consistent with the finding that near transfer problems were most effected, also seen in Level 6.

7.4 Discussion

When considering the right amount of help to offer students, there are competing reasons to provide more or less support. Students who receive more scaffolding may be unable to perform well when scaffolding is removed, but students without enough help may be unable to develop the skills at all. Vygotsky's concept of zone of proximal development recommends learning to happen in areas where students are unable to do a task alone, but are able to complete it with help, but it does not give much guidance in how to find those concepts or skills. In terms of our mastery homework system, this could be evaluated by how easy a level is to master; levels that are easy for all may not push students to think critically, while levels that have low mastery rates may indicate that more help is needed. This particular study showed that students who had higher mastery rates on their training level were able to perform significantly better on a more difficult level while spending significantly less time than those with the low mastery rate, suggesting that a mastery rate of 35% is below the ZPD.

Whether or not the 70% mastery rate of the group seeing extra scaffolding is ideal is not clear, but it is better than 35%. Within Level 6 itself, the greatest difference between groups was in the most similar problems, and achievement differences became less pronounced as transfer increased. This pattern was consistent in delayed feedback levels as well, though less significant.

This suggests that levels with low mastery rates are not necessarily creating spaces for students to self-motivate, but can possibly be limiting students' learning, particularly when following levels depend on the prior levels' skills. The effect of a single treatment was diluted as problems became more removed from the original training, but the basic skills still improved students' ability to perform the assessment at all, which was already a level of difficulty higher than the training level. The fact that students spent less time overall contradicts the idea that more time with material will necessarily result in better performance.

Looking back at Figure 7.1, we have many levels which have low mastery rates. The results from this study encouraged us to continue to investigate the effect of mastering on students' ability to demonstrate competence on a level's skills. In our next iterations of the class, specific low-scoring levels were be converted for half the class from mastery-style into immediate-feedback questions, and a study was done to see the effects of breaking down multi-step levels into smaller levels. This will reduce the amount of time spent by students on levels with heavy computation or multiple competencies, and we can evaluate if students' performance is affected in later levels and on bi-weekly quizzes. Our results here suggest that reducing strain on students due to very difficult mastery levels allows them to perform better, and set the groundwork to be able to test that further.

Chapter 8

Limits of Mastery: Grain size of levels

8.1 Introduction

A major component of mastery learning is breaking content into manageable pieces with clear competencies for students to master. Since the original implementation, each iteration of the course gives new opportunity to refine and improve the content and delivery of the mastery-style homework to better serve students. After the large restructuring of the delivery and content after its first year, the mastery homework still had many homework levels where the majority of students were unable to master; these levels continue to be a natural place to look for the limits of mastery-style learning, and this chapter will discuss another experiment meant to test those limits.

Vygostky's concept of the zone of proximal development (ZPD) encourages learning to take place between what students can and cannot do, but in skills that they can only do with help [32]. If students are unable to master, it is possible that the system is either not providing effective help, or that the material is not well-suited for mastery. In the previous implementation, it was discovered that students who received extra scaffolding in their problem statement of a difficult mastery level outperformed their classmates (who had not seen extra scaffolding) on a more difficult follow-up assessment (with no scaffolding), while still spending less time overall [136]. Especially because student frustration has historically been of special concern in the course, ensuring that the content is achievable is important.

A common complaint from students is mastery's condition to redo an entire set while only missing one question in the set. Repetition is inherent in mastery, but if the scope of a level is too broad, covering disparate topics, frustration may be warranted. There is also concern that creating levels that are very narrow may make problem solving too disjointed, and students may not see the parts of a problem as being a whole process. This study intends to examine the effects of different grain-size content levels of mastery-style homework on students' behavior and learning.[1]

[1]The work was published in the Physics Education Research Conference Proceedings in 2018 as "Effective Grain-Size of Mastery-Style Online Homework Levels" [143]. The experiment was designed and implemented by Noah Schroeder and analysis was done by the author.

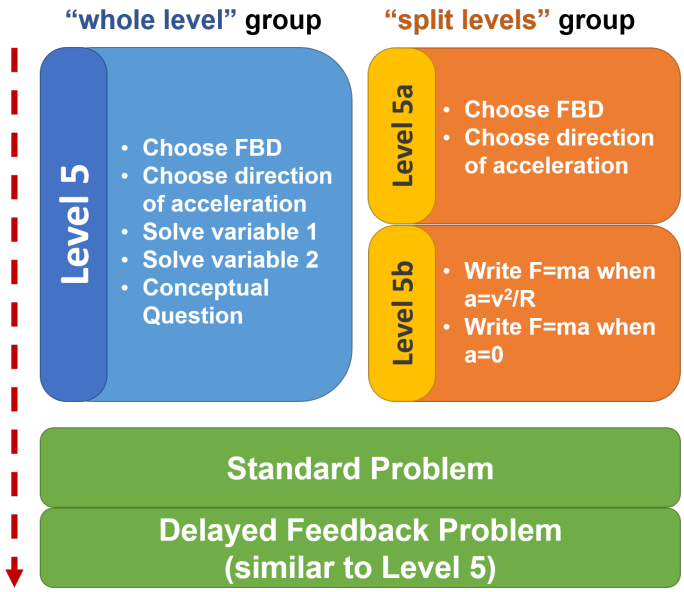


Figure 8.1: Structure of different treatments and following assessments for each group.

8.2 Methods

Like the previous experiment, the students in the course were split randomly between two online courses that were otherwise identical and each of the online courses received one of the two different treatments. A difficult level covering uniform circular motion was split into two smaller levels, with half the class receiving the "whole level" (the unaltered difficult level: Level 5) and the other half receiving two "split levels" (Levels 5a & 5b). By the random splitting, 195 students saw the "whole level" and 186 students saw the "split levels." The treatments were the last levels of Week 7, following four levels on friction. The friction levels have notoriously been difficult, but with the homework split between two bi-weekly assignments, students only had to complete one friction level immediately before the level(s) on uniform circular motion.

Following the different treatments, both groups saw a computational standard problem (immediate feedback with unlimited tries) and a delayed feedback problem (feedback given after the deadline) which was a version of the original "whole level" mastery. Thus, there were only three unique versions of level 5 (since the fourth version was removed to use as the delayed feedback assessment), so only three unique versions of levels 5a and 5b were created. The structure of the treatments and assessments for the two groups is shown in Figure 8.1; the figure also details how the tasks from the whole level were adapted to smaller levels. Note that the "whole level" group saw conceptual and computational problems that the "split levels" group did not.

8.3 Results

Comparing only the identical assessments (the standard problem and delayed feedback problem), students' performance was statistically equivalent for the two groups. Additionally, the average total time spent by a student in either group on the homework levels, standard exercise, and delayed feedback were also statistically

Table 8.1: Average scores for students with each treatment on common assessments, and total time spent on homework levels, standard problem, and delayed feedback. All are not significant differences with p >0.1.

	"whole level" group	**"split levels" group**
Standard Problem Average Score	$80.0 \pm 2.9\%$	$85.0 \pm 2.7\%$
Delayed Feedback Average Score	$56.9 \pm 3.9\%$	$56.6 \pm 3.9\%$
Average Total Time Spent	45.9 ± 2.6 min	42.6 ± 2.1 min

equivalent. Actual values for performance and overall time are shown in Table 1. The distribution of the total time across activities was also similar between the groups.

Within the homework levels themselves, students progressed through the smaller split levels quicker, with a higher rate of mastery overall (around 70% compared to 30%). Students' progression is shown in Figure 8.2. The mastery rate for level 5 is significantly different than level 5a and level 5b, with effect sizes of 0.8 and 1.0, respectively, and p <0.0001. To more fairly compare similar content, however, Figure 8.3 shows students' progressions through questions 1 and 2, which were identical question types in level 5 and level 5a. The group who practiced on the split levels had a higher rate of mastery on their second try by 24.7 ± 7.0%, corresponding to an effect size of 0.55 with p <0.0001. By students' third try, the split levels group had a 15.0 ± 6.3% higher rate of mastery on those two problems than their counterparts, corresponding to an effect size of 0.35 with p <0.005.

Although students spent similar amounts of time on the homework levels, it was possible to see how students were spending that time, then classify the time they spent as necessary or unnecessary practice. Specifically, unnecessary practice was defined to be time that students spent working on problems that they had answered correctly on a previous version and their current version, while necessary practice was defined to be time spent by students answering questions that they had not previously answered correctly. A scatter plot of students' time spent on necessary versus unnecessary practice is shown in Figure 8.4, where each data point corresponds to a student. One can see a higher concentration of students who saw the "whole level" in the top left quadrant, corresponding to large time spent on unnecessary practice and less time on necessary practice. Students who spent 16 or more minutes on unnecessary practice were capped and plotted at 16 minutes on the scatter plot, but the distributions ranged up to a maximum of about 40 minutes for students in the "whole level" group compared to 20 minutes for the "split levels" group. Because the plot makes it difficult to see overlapping values in the lower range of unnecessary time, data was also binned by taking a fraction of unnecessary time over the total time students spent on their homework levels. The distribution of students' percentages of unnecessary practice time is shown in a histogram in Figure 8.5. Both the average and median of students' unnecessary practice time are more than tripled for the students who saw the whole level compared to the split levels. The students who saw the "split levels" averaged 6.0 ± 0.5% of their time redoing previously mastered problems, compared to 18.7 ± 1.2% for students who worked through the "whole level" mastery. This is an effect size of 1.0 with p <0.0001.

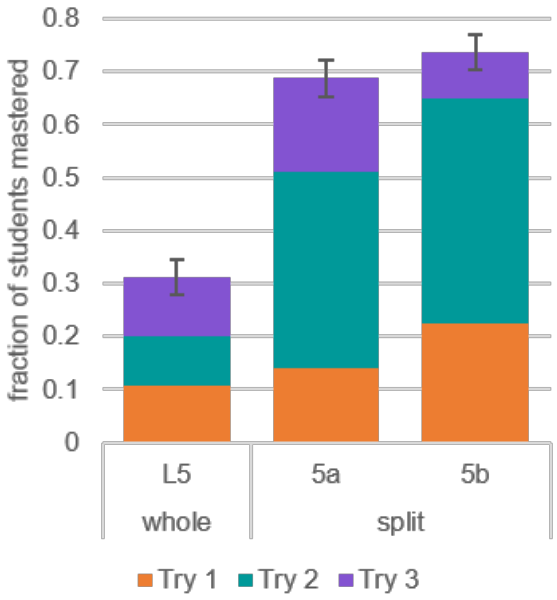


Figure 8.2: Mastery rate (fraction of students mastering) for whole and split level treatments, by try.

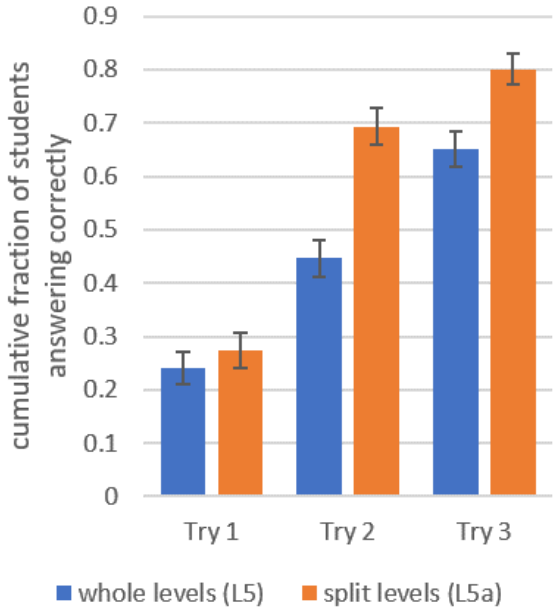


Figure 8.3: Student progression through similar problems, questions 1 and 2.

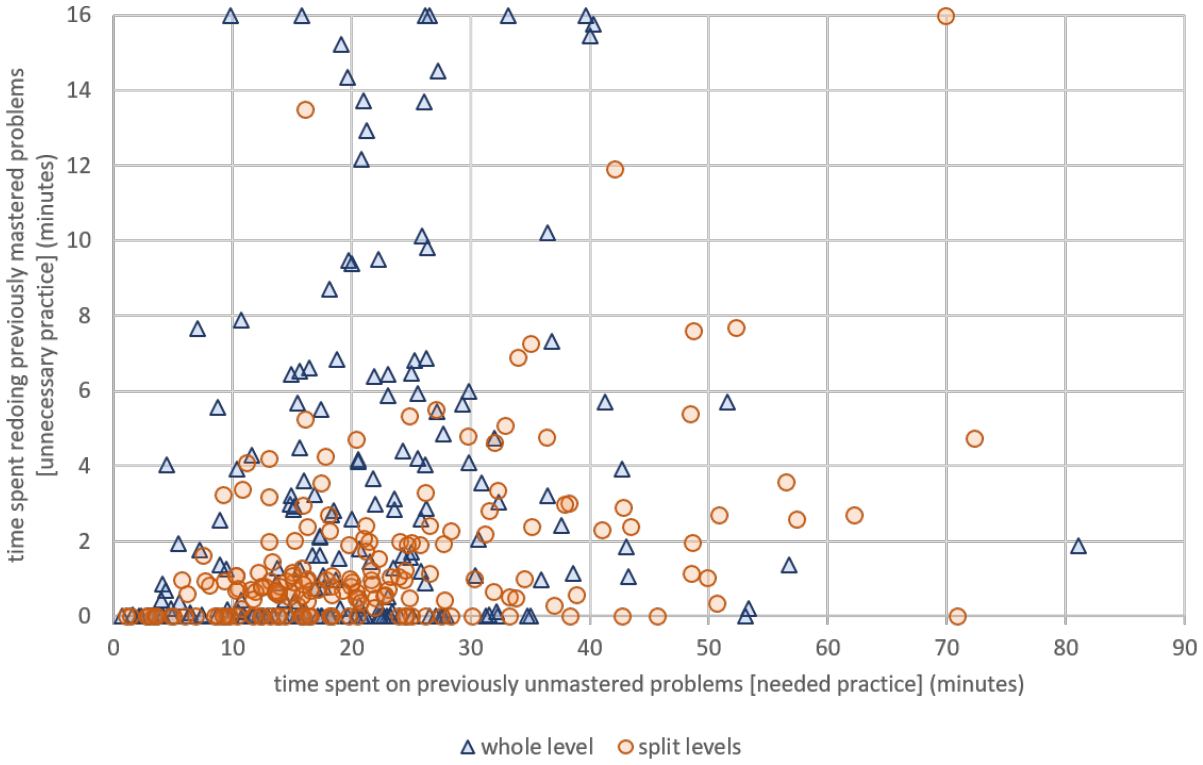


Figure 8.4: Scatter plot of students' time spent on necessary versus unnecessary time on the mastery levels. Students who spent 16 or more minutes were plotted at 16 minutes to make the plot more readable.

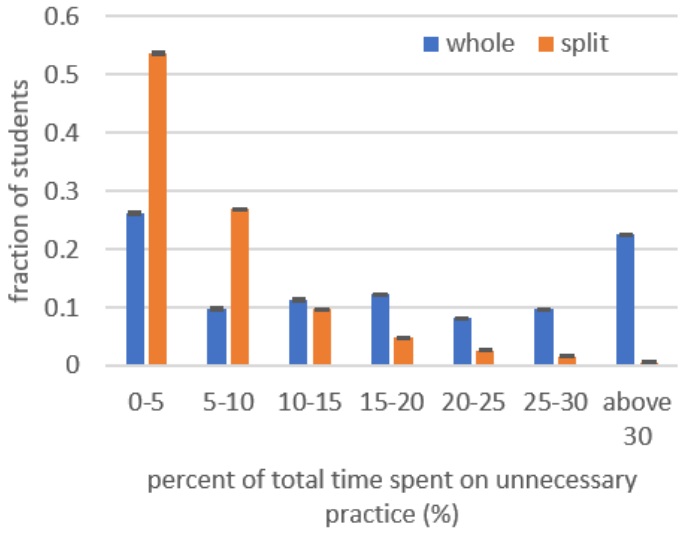


Figure 8.5: Histogram of percent time students spent on unnecessary practice out of total time spent.

8.4 Discussion

As mastery-style online homework continues to be refined at the University of Illinois, these results encourage further study into how changing the grain-size of levels will affect students' performance and frustration. Although there was no evidence that splitting a larger level into smaller pieces improved students' ability to solve problems, we also did not see that it significantly hurt their ability to do multiple-step computational problems. Smaller grain-size levels did, however, allow students to more often master their homework levels in fewer tries, which should help temper student frustration and bolster confidence. Additionally, students were able to prioritize their time more effectively to work on problems that they had not mastered on their previous or current version, wasting less time re-doing problems that they can already correctly answer.

These results are good preliminary evidence to probe the research question further for a more definitive result. In later implementations, the levels can be redone to be more similar in content. Particularly, the "whole level" group had the benefit of practicing computation and the conceptual questions, while the "split levels" group did not. These skills were tested on the assessments, which may have given the "whole level" group an advantage. In the next experiment, a third level can be added to the split levels which will ask students to practice computation in the context of circular motion, and the conceptual question could be removed from the single level and delayed feedback assessment. The assessments in general can be re-evaluated to include more fine-grained questions which can illuminate where students make mistakes, and how those mistakes correlate to their performance on the homework. Students can also be asked about their level of frustration; it is expected that more successfully mastering levels will make students feel more positively about the homework, but survey questions can give more direct information that was not gathered in the first iteration. The study can also be duplicated in a static friction level in the same course. The level is a good candidate for this treatment because it is historically difficult and the questions themselves have very different success rates. The variation in question performance suggests that the questions require different skills and that students have different levels of comfort with those different skills.

Moving forward with mastery-style online homework, it is imperative to students' morale for the content to be manageable, particularly for a group that is already at-risk to question their place in engineering. The current focus to improve our mastery-style homework is to mitigate very difficult levels; breaking content into smaller pieces may give students more opportunities to feel success and manage frustration. If doing so does not hurt overall performance, smaller grain-size levels may be a positive solution. Further study with more intentional assessments and an additional area to test the grain-size of the mastery delivery will provide more information about how to split content effectively to help students master content and feel success.

Chapter 9

Bi-weekly Quizzes with Retakes

9.1 Introduction

In addition to mastery-style homework, another attempt to provide support to Physics 100 students was the introduction of bi-weekly quizzes with retakes. The strategy was motivated by success in an introductory mechanical engineering course in decreasing DFW (Drop, Fail, Withdraw) rates using bi-weekly quizzes with retakes in conjunction with other classroom restructuring. The implementation of frequent testing was a large part of the restructure and we hoped that providing students consistent and low-stakes feedback via testing would be effective for Physics 100 students as well.

Mastery learning is a type of frequent testing, and thus, many of the theoretical advantages of mastery are also relevant for frequent testing as a whole. Using frequent testing takes advantage of the testing effect, which researchers have documented experimentally in multiple studies as aiding performance and retention [23–28]. Frequent testing also gives formative assessment to students and instructors about students' competencies and areas for growth to improve student learning [18]. Giving students more frequent feedback to shape their study provides opportunities to improve in areas that need the most work and demonstrate that improvement on the retakes.[1]

Education literature has focused on types and frequency of feedback enough that meta-analyses exist of studies which evaluate feedback [58, 144, 145] and specifically timing and frequency of feedback [146–148]. Although many cite feedback interventions as positive, there is a lot of variation in their effectiveness, suggesting that students' responses are sensitive to the implementation of the feedback [58, 144]. A synthesis of meta-analyses across many learning factors by John Hattie remarked on the most effective forms of feedback [58, 149]; he saw the largest effect sizes improving performance when feedback provided cues to students, used audio, visual, or computer-based assistance, and he saw little improvement motivated by extrinsic rewards. In their meta-analysis, Kluger and DeNisi saw that the largest gains in feedback studies which gave students specific goals for challenging tasks [144]. They also note that effects are larger when students perceived a low level of threat to their self-esteem, citing the extra attention students can pay to the feedback when not sharing attention with the threat [58, 144].

[1] More elaboration on the testing effect and formative assessment are included in Chapter 2.

This is consistent for students who suffer from test anxiety, a phenomenon first coined in 1952 [150]; cognitive load can be taken up by "task-irrelevant responses" which interfere with or detract from performing the task for the exam [150–152]. Mandler and Sarason characterize the irrelevant responses with examples: "feelings of inadequacy, helplessness, heightened somatic reaction, anticipations of punishment or loss of status and esteem, and implicit attempts to leave the testing situation" [150]. People belonging to marginalized communities often have higher levels of testing anxiety; Hembree's meta-analysis showed that Black and Hispanic students tended to have more test anxiety than white students and female students consistently had more test anxiety than their male classmates [152]. Considering that Physics 100 tends to represent slightly higher fractions of women, Black, and Hispanic students than the engineering college as a whole, it is important to be conscientious of this risk. However, multiple meta-analyses discovered that performance for students with high test anxiety was improved with more frequent testing [152, 153]. Sulan et al saw in their study for music students that students who believed that frequent testing was hurting their performance actually saw the highest gains from more frequent testing, which emphasizes that students' affective response is not analogous with their performance [154]. However, it is still important to consider the negative effects on students' morale, regardless of performance outcome, from a moral standpoint and in light of the adverse effect of student frustration in our initial mastery implementation [11]. In our course, the addition of bi-weekly quizzes and retakes was intended to create a low stakes environment for students to receive frequent formative feedback, ideally with the benefits of the testing effect and feedback outweighing the negative effects of anxiety around testing.

Specifically in the realm of frequent testing, educational psychology researchers have been studying its effectiveness since the 1920s [148]. In 1923, Jones dedicated his dissertation to the effects of brief examinations, finding that students performed better with frequent, brief testing than those who did not have additional testing [24]. Two studies from the 1930s showed that by giving additional testing to low-performing students, the students performed equally well as their classmates on their final exam [155, 156]. A study by Mawhinney et al further showed that students' self reported study time was more consistent and increased as testing frequency increased [157]. Since then, results from frequent testing studies have varied in effect sizes, though most show positive effects for testing [146, 148].

A meta-analysis by Bangert-Drowns saw that the effect sizes were largest when measuring against control groups which saw no testing, and that effect sizes increase with the frequency of testing but have diminishing returns. In particular, when considering studies which looked specifically at different amounts of testing within the same populations of students, effect sizes averaged 0.49 for high frequency testing and 0.23 for medium frequency, which corresponded to tripling the number of exams for double the effect. From his analysis, he created a model for expected effect sizes for improved performance over students with no previous testing, depending on the frequency of additional testing, which is shown in Figure 9.1 [146]. His model predicts that effect sizes will increase proportionally to the fourth root of the number of tests, and a fit for his curve is approximately defined mathematically as $ES \approx 0.35\sqrt[4]{N}$, where ES is the effect size and N is the number of tests. A meta-analysis by Başol and Johanson contradicted Bangert-Drowns findings, citing that they saw no correlation between the frequency of testing and effect sizes, though their categorization for low to high frequency testing was ambitious [148]. They cited only daily or every other day testing as high, weekly exams as medium frequency, and every other week or fewer as low frequency [148]. In comparison, Bangert-Drowns found the average of frequency for medium frequency testing averaged 7 exams in 15 weeks [146], which corresponds to low frequency for Başol and Johanson. Using Bangert-Drowns' diminishing

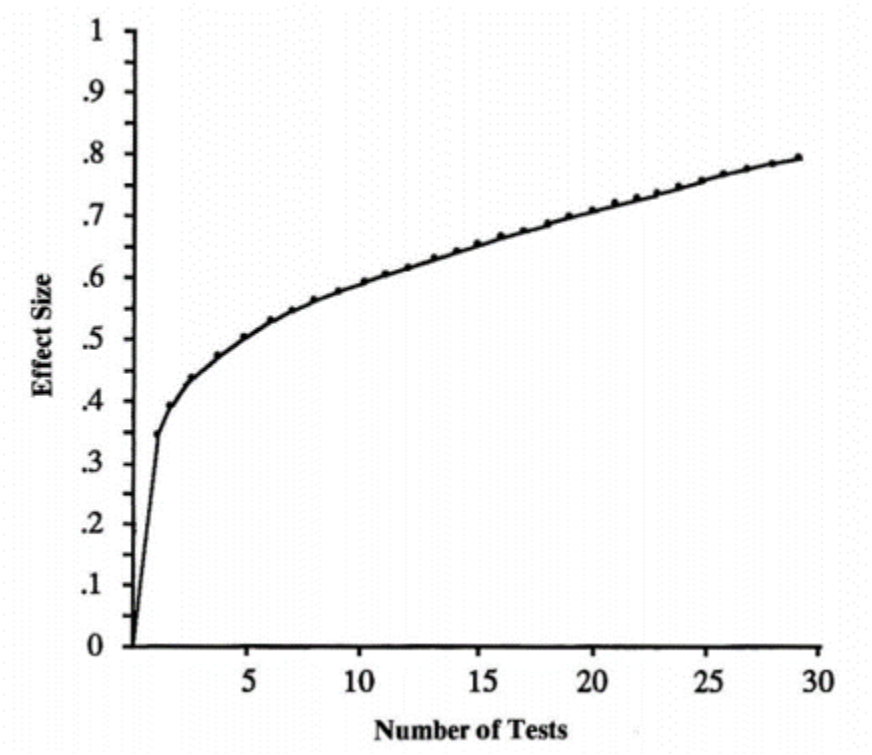


Figure 9.1: Bangert-Drowns' model for expected effect sizes on exams when increasing test frequency from no intermediate testing [146].

returns model, Başol and Johanson's classifications would not likely be able to distinguish between their categories, as the largest effect size differences happen all within their "low" frequency testing category. In physics 100, quizzes were given every other week, with the possibility of retakes on off weeks, placing our testing frequency in the medium to high range.

9.2 Methods

The addition of bi-weekly quizzes was done for all students in Physics 100 in 2017. Due to the addition of these quizzes, the structure and grading of the course was altered to accomodate the extra work for students; as of 2017, the breakup of activities described in Chapter 1 was changed. Online quizzes were removed and the distribution of grading was changed. The bi-weekly quizzes were assigned 20% of students' grades, which was redistributed from the 15% that was previously the online quizzes, and 5% from the final exam (The final was changed to be worth 15% instead of 20%.). Table 9.1 shows the updated grade distribution and timeline for students, which reflects these changes and includes the breakup of the homework into two smaller assignments from 2015.

Students had access to their bi-weekly quizzes from Friday through Monday, and could schedule their quizzes during that time period. This was made possible by proctoring the quizzes through PrairieLearn at the Computer Based Testing Facility (CBTF) at the University of Illinois, which has previously shown improved grades for students compared to paper testing for computer science students [158]. Physics 100 students

Table 9.1: Activities and their timing for a typical unit of content in Physics 100, reflecting changes in homework timing from 2015 and relevant grade re-distribution for bi-weekly quizzes in 2017. The last 30% of students grades come from their midterm (15%) and final exam (15%).

Day	**Activity**	**Portion of Grade**
Before Friday	Prelecture and Checkpoint	9.5 %
Friday	Lecture	4.5 %
Monday	Office Hours	–
Sunday, 11pm	Homework A Due	15 %
Tuesday, 8am	Homework B Due	
Tuesday	Discussion Sections	20 %
Wednesday		
Thursday		
Friday - Monday	Quiz Available	20 %

made appointments through the testing facility's online interface and would show up at their assigned time, where they could take the quiz while being proctored by student employees. Although the "online quizzes" that were removed from the course were also called quizzes, these new bi-weekly quizzes more traditionally represented a testing environment. Their online quizzes had no time limit (as long as they finished before the deadline) and students were allowed any notes or resources to complete them. In contrast, the testing facility does not allow students to bring notes, and the bi-weekly quizzes were timed and 50 minutes long. The PrairieLearn software at the testing center displays students' scores after submitting each question rather than finishing the quiz and getting their score at the end. This was not ideal; Butler and Roediger showed results that student recall is better if feedback is given at the end of a quiz rather than question by question [100], but this factor was not in our control. Particularly for students with test anxiety, seeing points lost in real time can be stressful and distracting.

The material for the course was divided among 6 quizzes, which were given staggered in weeks after the students had completed the other learning activities for that content, the schedule is shown in Table 9.2. The first five quizzes were focused on specific content, but still contained some review questions from previous weeks, and the last quiz was a review from all weeks of content to prepare students for the final exam. There were about 15 questions on each of the quizzes, which were a combination of old exam questions from Physics 211 (the target course of the preparation in 100) and similar problems to the mastery-style homework. The questions from previous exams were chosen intentionally to be questions which students in Physics 211 scored above 70% on their exam. Some questions were used as assessments for another study and were significantly harder than we would have otherwise included, but the quizzes as a whole were intended to be achievable.

The PrairieLearn platform allowed us to create multiple versions of questions and randomly select from among multiple versions to mix and match for each student. Especially because the quizzes were self-scheduled over many days, extra versions of questions were used to keep students who took the quiz early in the window from passing on questions to those who took it later on. Chen et al showed that with only one version of questions, "cheaters" tended to have a score inflation of about 13%, but that dropped to about 3% if the pool for each type of question increased to 2, and 2% when increased to a pool of 4 [159]. He

Table 9.2: Schedule of Bi-Weekly Quizzes in relation to the material being learned. Recall that only the first 8 weeks of the course are new content-based instruction, and the second half of the course focuses on problem solving.

Material Learning	**Quiz**	**Material on Quiz**
Week 1	-	-
Week 2	Quiz 1	Week 1
Week 3	Quiz 1 Retake	
Week 4	Quiz 2	Weeks 2 & 3
Week 5	Quiz 2 Retake	
Week 6	Quiz 3	Weeks 4 & 5
Week 7	Quiz 3 Retake	
Week 8	Quiz 4	Weeks 6 & 7
(Week 9)	Quiz 4 Retake	
(Week 10)	Quiz 5	Week 8
(Week 11)	Quiz 5 Retake	
(Week 12)	Quiz 6	Review Weeks 1-8
(Week 13)	Quiz 6 Retake	

defined "cheaters" as students who studied the practice exam problems which mirrored the test questions at a disproportionate rate to their other studying, if it occurred during the time period after the exams were open and before their exam. Their practice and actual exams were all through PrairieLearn in his study, making this insight especially relevant. Most of our quizzes had 4-5 versions of each question, but some of the later versions of retakes had only two due to time constraints in creating the questions.

All students were expected to take the original quiz, and the retake was optional. Students' final quiz score was weighted as 1/3 of their original quiz score and 2/3 of their best score. This meant that if students performed less well on their retake, they did not risk losing points and just received their original score, but students who performed better on their retake had the ability to raise their score significantly.

As a measure for the effectiveness of the quizzes, both the midterm and final were repeated from Fall 2015. Variation between students year to year is historically not significant, so this was done to allow us to compare exam scores between students who used the bi-weekly quizzes (2017) and those who did not (2015). Comparison of these populations by ACT scores will be addressed in results to give better context of their comparability.

9.3 Results

9.3.1 Quiz and Retake Behavior and Performance

Students' performance on the quizzes was concernedly lower than we had anticipated, particularly as the semester progressed. Students' retakes scores tended to be similar to the original average of the quiz,

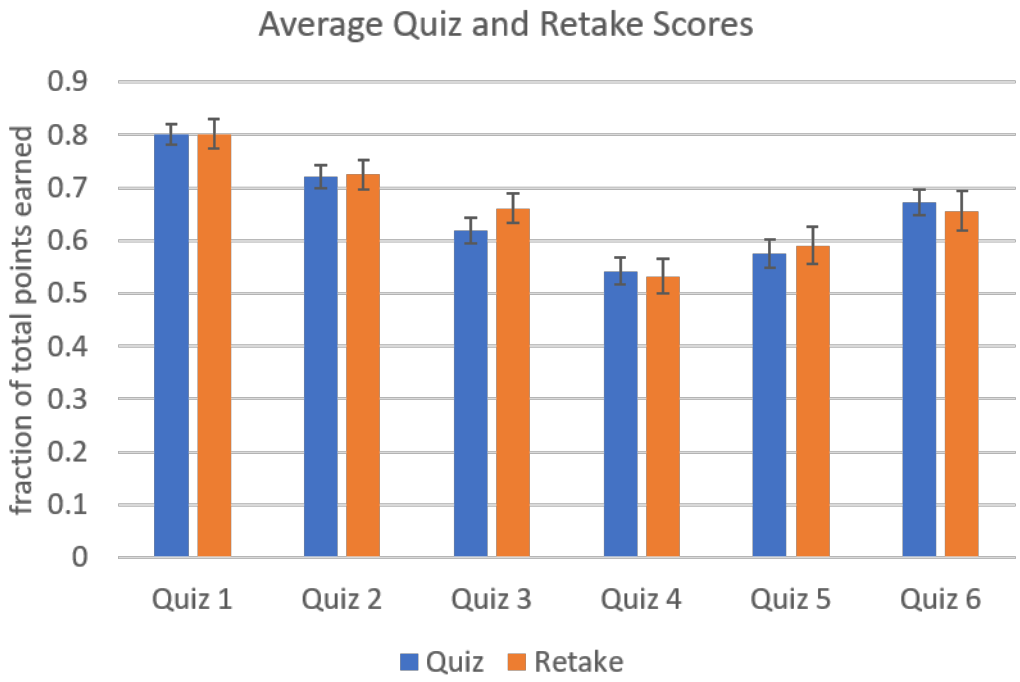


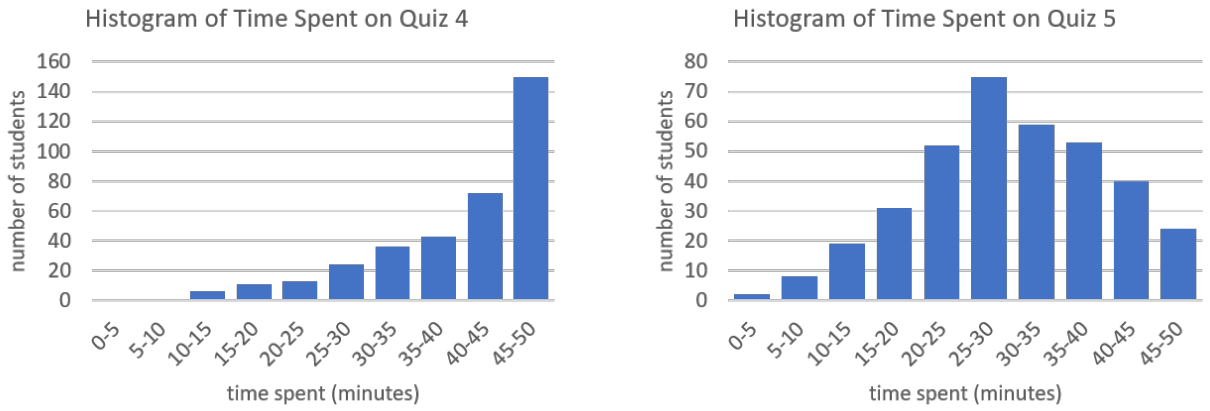
Figure 9.2: Average scores for each quiz and associated retake.

suggesting that students who performed poorly were able to use the retakes to meet their classmates' scores. A plot showing students' average quiz and retake scores is given in Figure 9.2.

After Quiz 4, the quizzes were adjusted to be only 12 questions (instead of 15 or 16) in light of students' declining scores with increasing content difficulty. The distribution of time spent on Quiz 4 had over 40% of students using between 45-50 minutes (with the time limit of the quizzes being 50 minutes), suggesting that many students were not getting enough time to complete all the problems. By making this change, we were able to successfully reduce the portion of students spending over 45 minutes to about 7%. The histograms of time distribution for comparison from Quiz 4 and 5 are shown in Figure 9.3.

Students who used the retakes tended to be lower scorers, so a more accurate comparison to show improvement by the retakes is shown in Figure 9.4, which shows the original quiz and retake scores only for students who took both. On this plot, one can see that the retake scores were higher for every quiz, but significantly so on only Quizzes 1-3. Over all quizzes, students who used retakes gained an average of 7% from their original quiz score to their retake (59% to 66% average scores).

On average for each quiz, there were only about 10-15 students (of around 400) who did not take the original quiz but only took the retake. Across all quizzes, an average of 57% of students used each retake, with the portion of students using the retakes ranging from about 45 - 70%, depending on the quiz. There was some correlation between the quiz scores and the fraction of students using the retakes, which can be seen in Figure 9.5. Quizzes 3-5 had the lowest average scores and the highest fractions of retakes happened in Quizzes 3 and 4. It is worth noting that Quiz 5 was given immediately after the midterm, which may be a reason that the fraction of retakes was low for its difficulty. To untangle this further, the fraction of students using retakes based on their associated quiz score (integrated over all quizzes) is shown as a binned histogram in



(a) (b)

Figure 9.3: Histograms of time spent on the 50 minute quiz for (a) Quiz 4 and (b) Quiz 5.

Figure 9.6. This plot suggests that about 70% of students used retakes if their score was under 70% on the original quiz, but even 40% of students scoring in the 80-90% range were using retakes.

By the end of the semester, students' final summative quiz scores (including the inflation of scores due to retakes) were primarily between 60 and 90%, as seen in Figure 9.7. About 23% of students scored below 60% on their quiz scores, which was a higher fraction than we anticipated or hoped, but more than half the class was able to score above 70%.

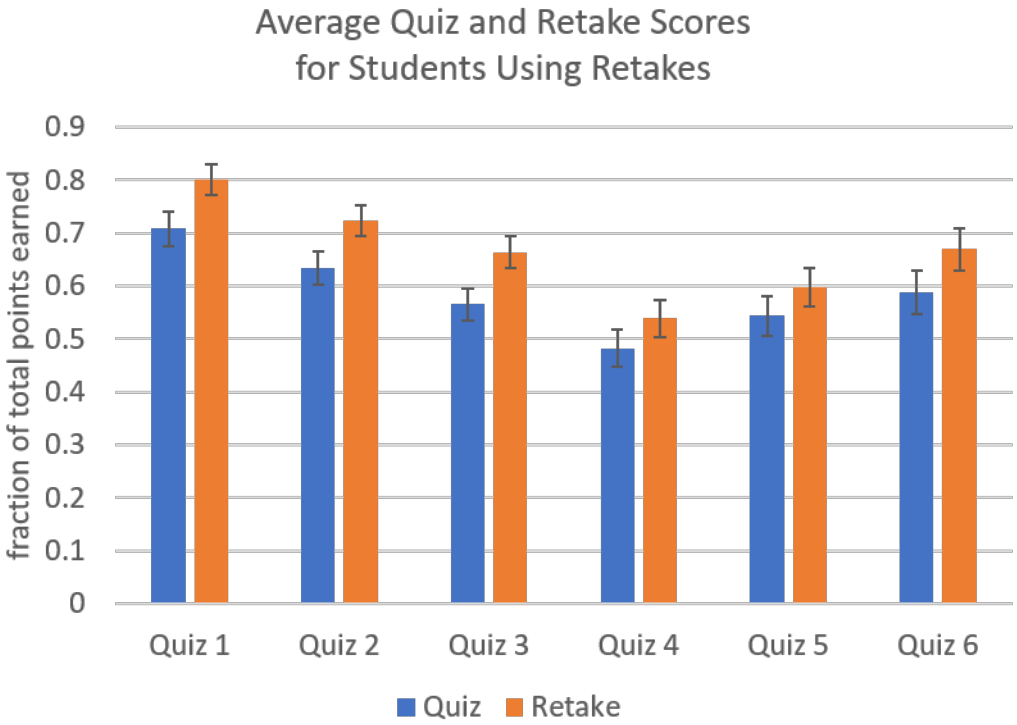


Figure 9.4: Average scores for each quiz and associated retake, including only students who took both the original quiz and the retake.

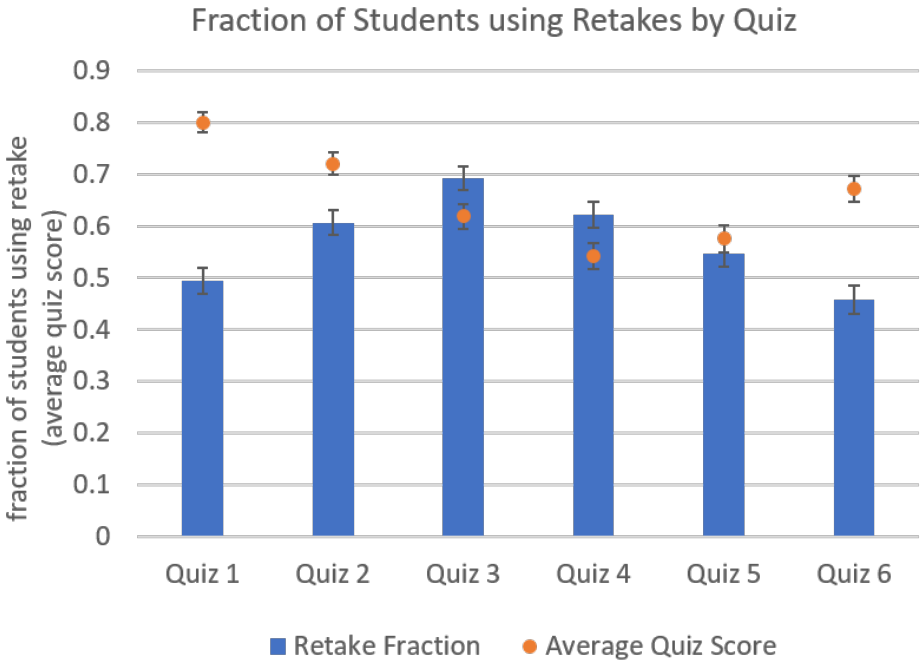


Figure 9.5: The fraction of students using the retake quiz (blue bars) and the associated average score for the quiz (orange scatter points).

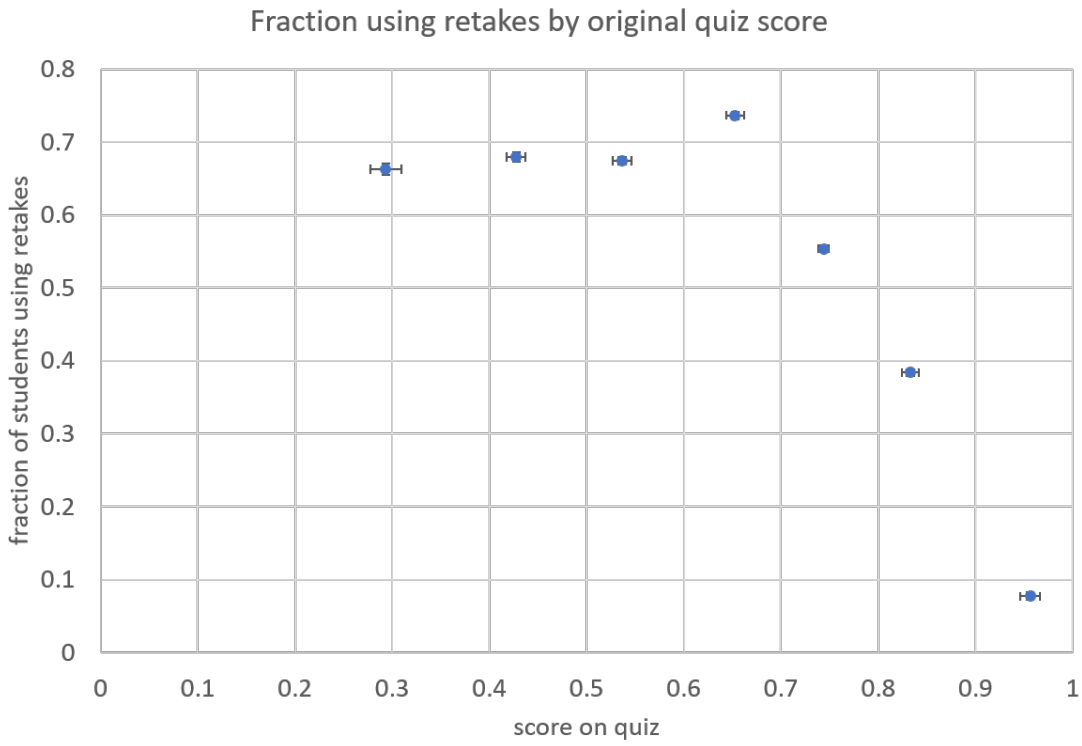


Figure 9.6: The fraction of students using retakes, binned by students' original quiz score integrated over all quizzes.

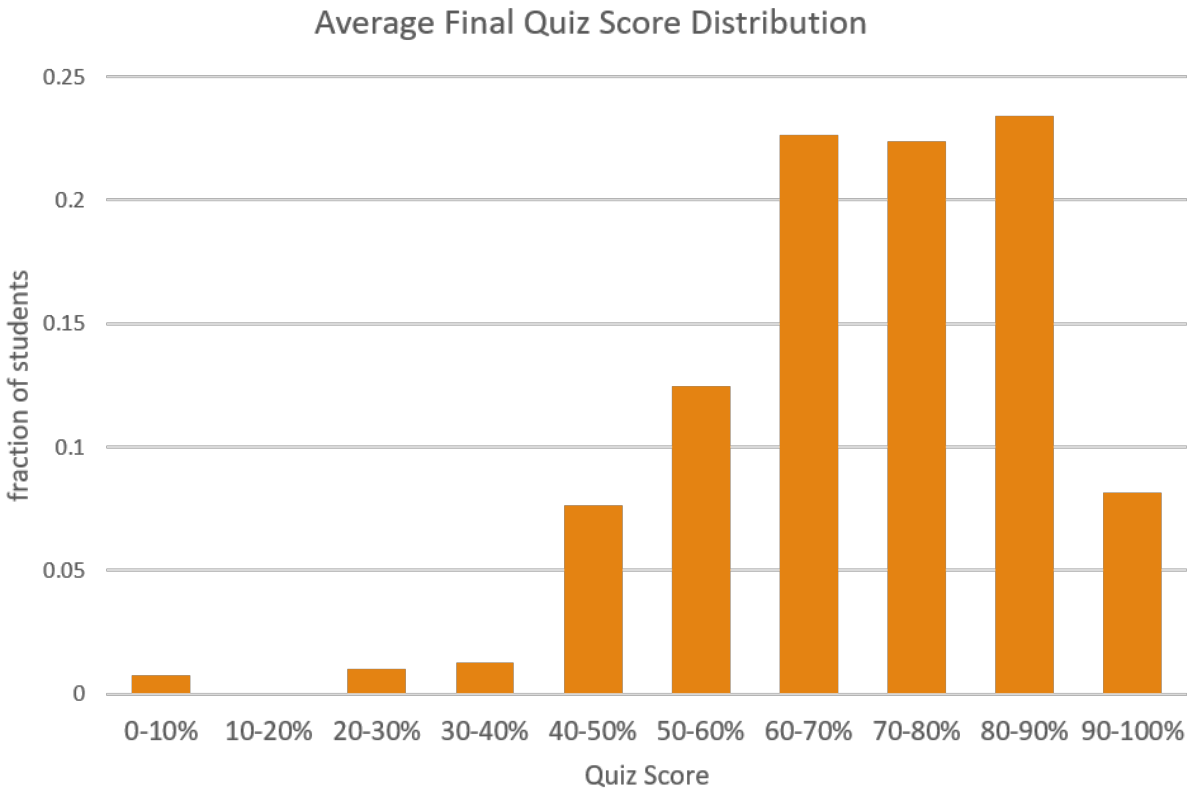


Figure 9.7: Distribution of students' final scores for quizzes at the end of the semester.

9.3.2 Midterm and Final Exam Performance

To gauge the effect of bi-weekly quizzes in Physics 100 compared to students in the course without the quizzes, students took identical midterms and final exams in 2015 (no quizzes) and 2017 (the introduction of bi-weekly quizzes). Historically, cohorts of Physics 100 students tend to be similar from year to year, and Figure 9.8 shows average ACT and physics diagnostic scores for both groups of students; based on these metrics, the two classes of students are not significantly different.

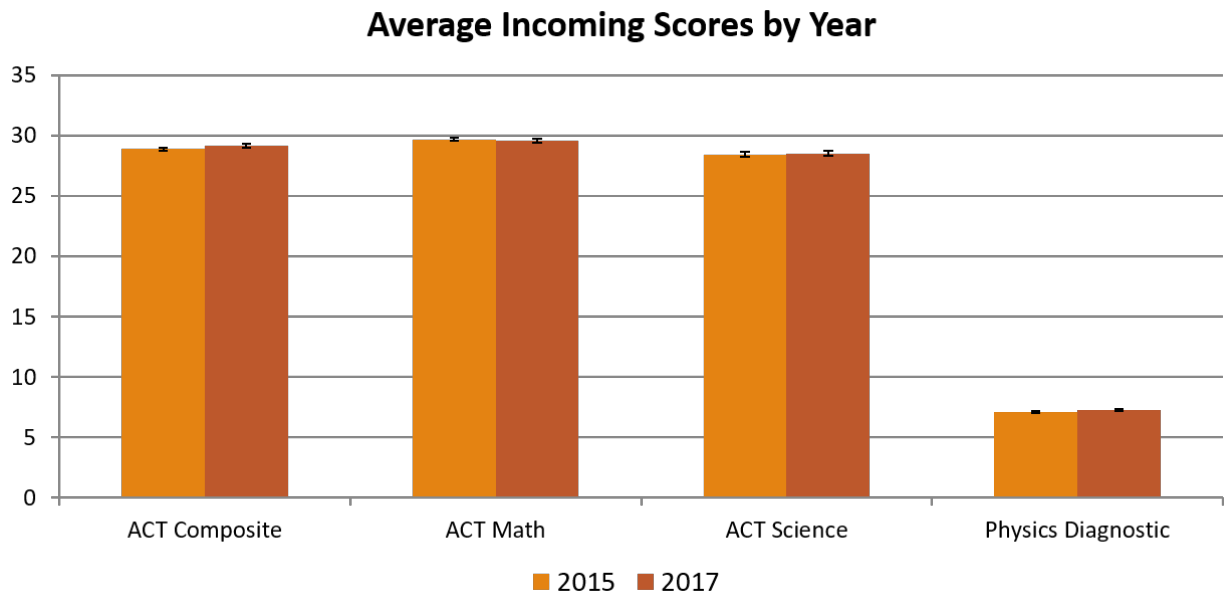


Figure 9.8: Comparison of students' incoming ACT and physics diagnostic scores for 2015 and 2017. The ACT scores are out of 36, and the physics diagnostic score is out of 13.

Midterm Exam

Despite their similarity coming into the course, the students in 2017 significantly outperformed their classmates from 2015 on the identical midterm. The average midterm score in 2017, with bi-weekly quizzes, was 68.2% compared to 61.6% in 2015. This 7% shift in the average score corresponds to an effect size of 0.35 and is significant with $p < 0.0001$. Based on bootstrap analysis of student scores, the 95% confidence interval for this effect size is between 0.27 and 0.45; the model by Bangert-Drowns [146] predicts an effect size 0.41 (for an additional 4 quizzes), which falls in our data's confidence interval.

Broken down by question, Figure 9.9 shows students' performance in both years; most questions saw a statistically significant improvement from 2015 to 2017. Looking at the distribution of scores, shown in Figure 9.10, there are significantly more students in each bin scoring above 70% in 2017. The midterm has historically been difficult for students, and Figure 9.11 shows the portion of students scoring below 60% for all recent years which used the same midterm. Since 2012, there were other changes to the course (such as the introduction and improvement of mastery-style homework), but conservatively between 2015 and 2017, the portion of students scoring below 60% decreased from about 44% to 30%, reducing by nearly a third.

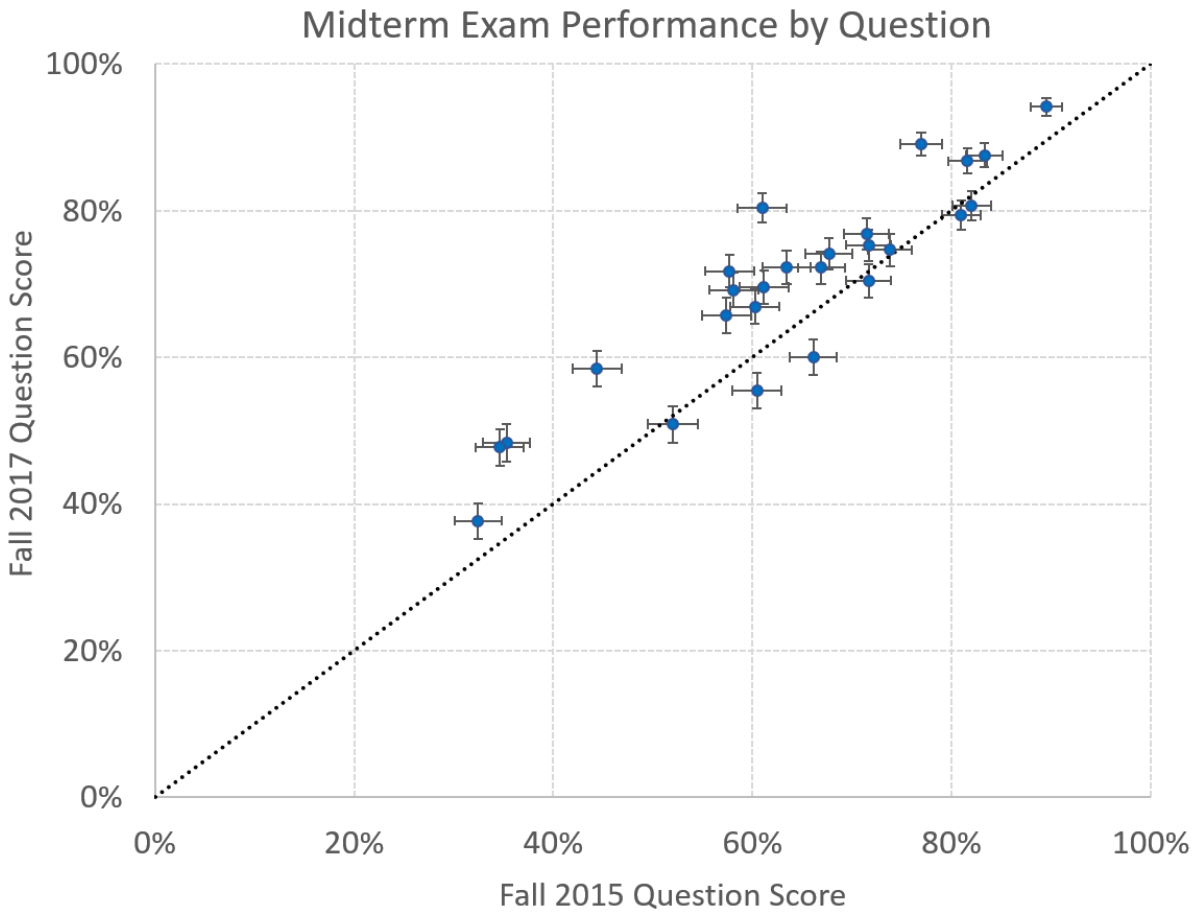


Figure 9.9: Midterm exam performance by question, with scores in the horizontal axis corresponding to 2015 and scores on the vertical axis corresponding to 2017. Points above the line indicate improvement in 2017 over 2015.

This reduction is an effect size of 0.56, statistically significant with $p < 0.0001$.

These improvements to students' scores were seen across student ability. By binning students according to their ACT Math score, as in Figure 9.12, there is improvement in each bin for students in 2017 over 2015, with scores converging for the highest performing students, most likely due to a ceiling effect.

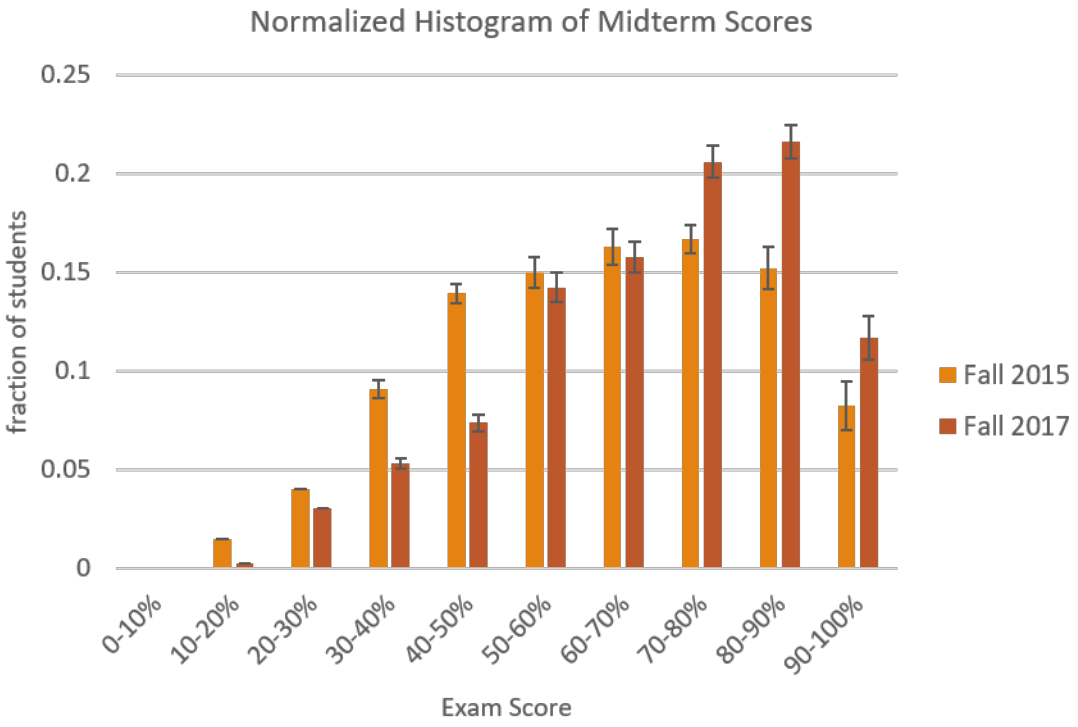


Figure 9.10: Distribution of students' scores on the midterm for 2015 and 2017.

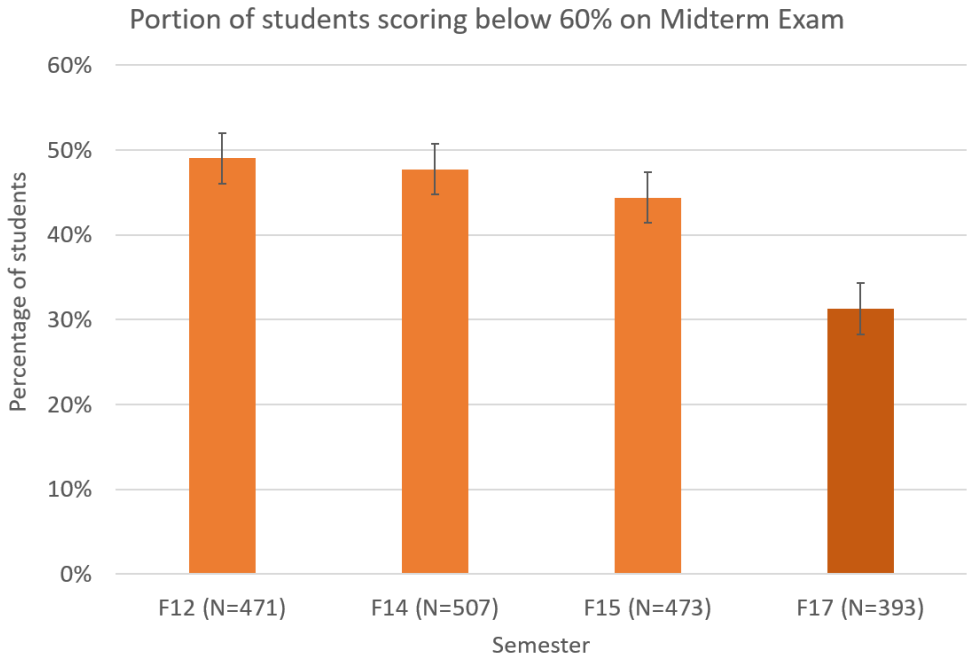


Figure 9.11: Percentage of students in 2012, 2014, 2015, and 2017 who scored below 60% on the midterm exam. The same midterm exam was used for all of these years.

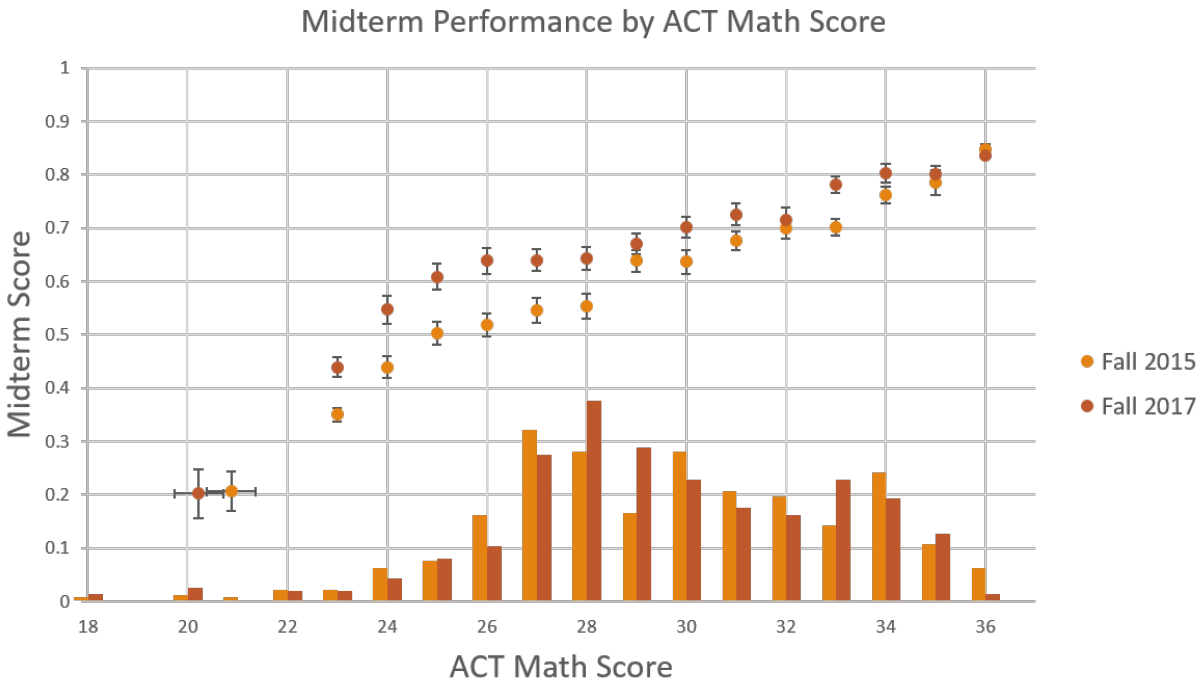


Figure 9.12: Midterm scores by incoming Math ACT score. The bars on the bottom represent a normalized histogram of the fraction of students in each bin. For ease of view, the histogram is not to scale with the vertical axis; the largest bar corresponds to a fraction of 0.16.

Final Exam

The same analysis was also done for students' performance on the final exam, which showed a smaller improvement than the midterm. Overall, students' average scores in 2015 and 2017 were respectively 62% and 65%, which was not a statistically significant improvement. Figure 9.13 shows that there were still some questions which showed improvement from 2015 to 2017, but less so than on the midterm. In 2015, students taking the final had already taken one exam (the midterm exam), so one would expect this result to be less significant by the Bangert-Drowns [146] model, as the largest effect comes from no tests to one test.

The distribution of student scores on the final is given in Figure 9.7; again, there was a reduction in the fraction of students receiving a score below 60%, this time from 46% in 2015 to 42% in 2017. This is a smaller reduction than seen on the midterm, with an effect size of 0.2, with a p-value of $p < 0.05$. Breaking up students by their incoming Math ACT score is also less clean; there are some gains for lower performing students, though they are not significant. The plot comparing scores in 2015 to 2017 by ACT Math scores is in Figure 9.15.

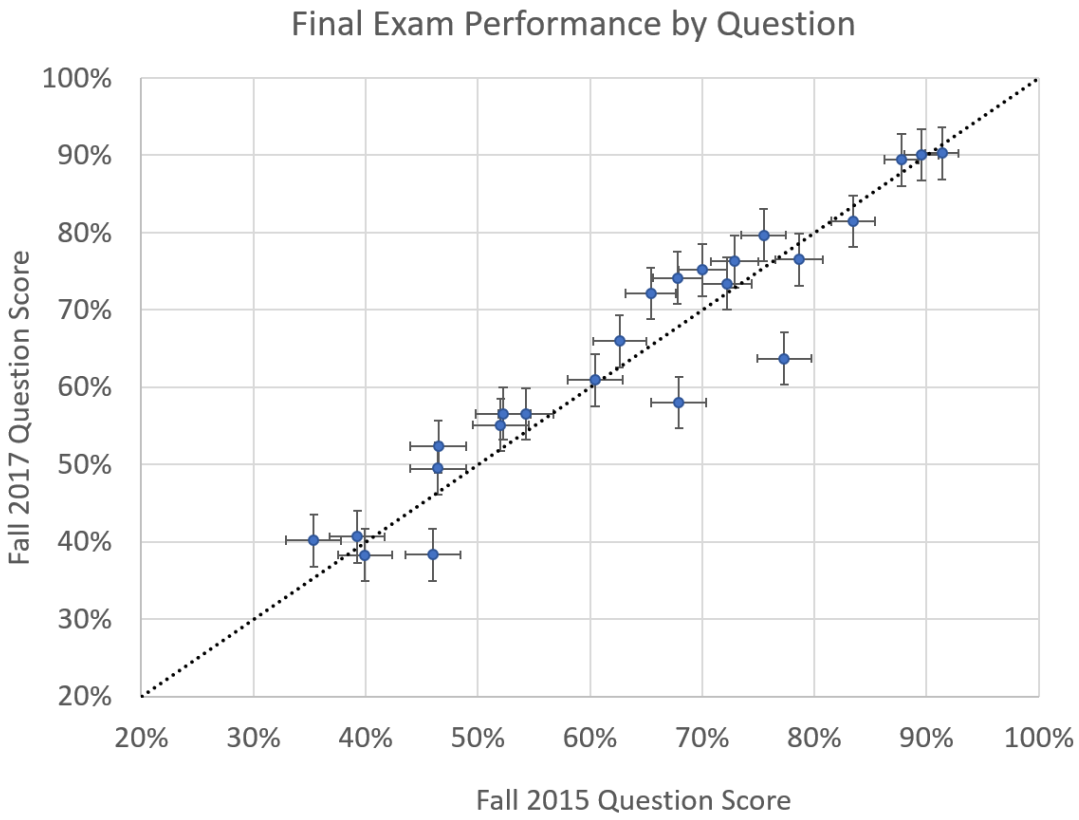


Figure 9.13: Final exam performance by question, with scores in the horizontal axis corresponding to 2015 and scores on the vertical axis corresponding to 2017. Points above the line indicate improvement in 2017 over 2015.

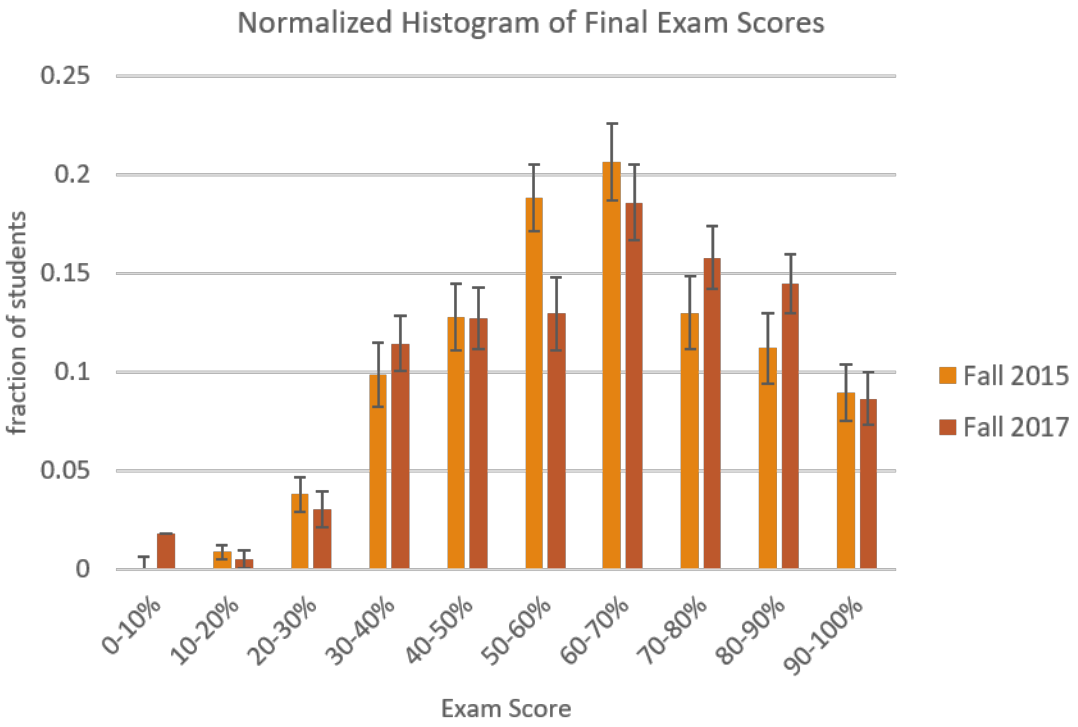


Figure 9.14: Distribution of students' scores on the final exam for 2015 and 2017.

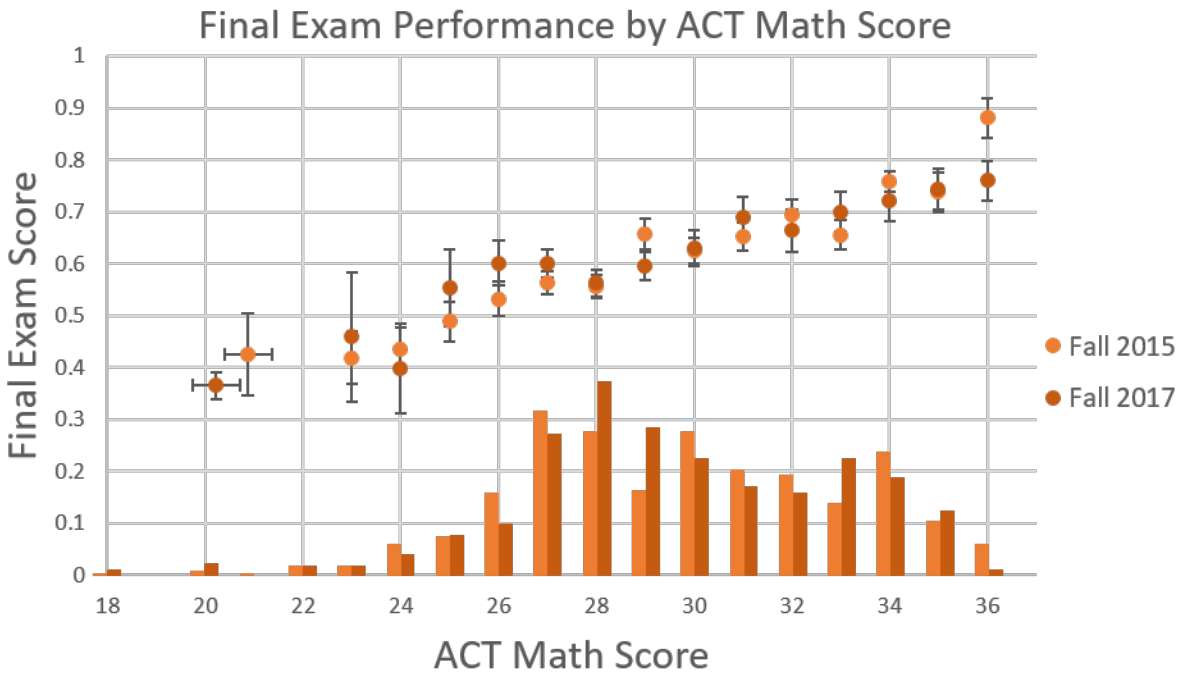


Figure 9.15: Final exam scores by incoming Math ACT score. The bars on the bottom represent a normalized histogram of the fraction of students in each bin. For ease of view, the histogram is not to scale with the vertical axis; the largest bar corresponds to a fraction of 0.16.

9.4 Discussion

The addition of bi-weekly quizzes to Physics 100 effectively improved students' performance on the midterm, including a statistically significant reduction to the fraction of failing students. On the final exam, the positive effects persisted, but not significantly. The effect size of the improvement for the midterm was within error of the predicted effect size from Bangert-Drowns' model [146], but the effect size for the final was lower than the model predicted. Even with the intervening midterm in 2015, the model predicts a gain of 0.2 standard deviations as the difference between having 1 intervening test versus 7, which were the treatments in 2015 and 2017. We saw essentially no improvement in average performance on the final, but still saw a decrease in the fraction of students who failed the exam. When considering the metric of failing students on the exam (scoring below 60%), the effect sizes are still well predicted by the model, with effect sizes of 0.56 (predicted 0.41) and 0.2 (predicted 0.2).

On the quizzes themselves, students who used the retakes were able to bring their scores up to their classmates' scores who did not use the retakes, suggesting that the option to retest was effective. However, on the later, more difficult quizzes, the original and retake scores were within error of each other; it is unclear whether students who performed poorly on the original quiz were low statistical fluctuations which were improved on a second attempt, or if students' improved scores were due to further study. Because there was no penalty for scoring lower on the second attempt, students had incentive to retake whether or not they had time to prepare more between the quiz and retake. In some cases, students who lost too many points early in a retake to outperform their original score abandoned the retake partly through; this was suggested by students' "good" performance early on in a quiz and all wrong answers for the later questions. Recall that students were given their scores in real time as they completed each question, so they were aware when it was no longer possible for them to do better. This may have also deflated retake scores.

Most of the quizzes had no practice quiz, which was a point of concern for many students. This was in part because the quizzes were used as an assessment for another study, and we wanted the quizzes to be sensitive to changes in the students' homework for half the class (which was being tested), rather than seeing a ceiling effect from studying the specific types of questions on the quiz. This analysis made no partition of students based on their different homework treatments, and the changes were made alternating between to the two groups of students week by week, so we expect no effect from that study on these results. In the next year, providing students access to pretests was a priority.

Overall, the addition of bi-weekly quizzes was effective at raising student midterm performance and reducing the fraction of failing students on the midterm, but the results were diminishing into the final exam. The retakes allowed students to improve their quiz scores, but having no penalty at all for performing poorly on the retake may have limited their incentive to use them conscientiously. The lack of pretests was a major complaint by students and their addition may encourage more effective study moving forward. We believe that the quizzes were valuable but that students' potential to use them effectively can be amplified.

Chapter 10

Retake and Practice Exams in Physics 211

10.1 Introduction

As a followup to the quiz retakes in Physics 100, retakes were added to the exams for the following semester of Physics 211, the target course. The motivation for their introduction was to ease the transition for Physics 100 students as they continued the engineering sequence; after having retakes for their quizzes in Physics 100, we wanted students to continue to have that space for growth in the introductory mechanics course.

The literature and theoretical framing to support retakes as a form of more frequent assessment are consistent with the previous chapter (Chapter 9). Allowing retakes specifically has been shown to improve students' scores across multiple subjects [160–164]. As one might expect, the effects of retesting are greatest if the retest is identical to the original test, but parallel retakes also show significant score improvements [162, 163]. Geving cites a meta-analysis by Kulik et al on practice exams as a proxy to predict an effect size of 0.2 standard deviations if the retake is a parallel version of the original exam, compared to 0.4 for repeated exams [162, 163]. Our exam retake implementation was similar to the quiz retakes; the retakes were parallel exams rather than identical exams.

The meta-analysis by Kulik et al on practice exams saw that students who studied more practice exams consistently saw greater effect sizes, without the diminishing return which we saw from frequent exams [163]. Figure 10.1 shows their model for effect sizes, both for exams which were identical and parallel to the practice exams. This suggests that providing practice exams and students' engagement with the practice exams will also affect our results on exams and retakes.

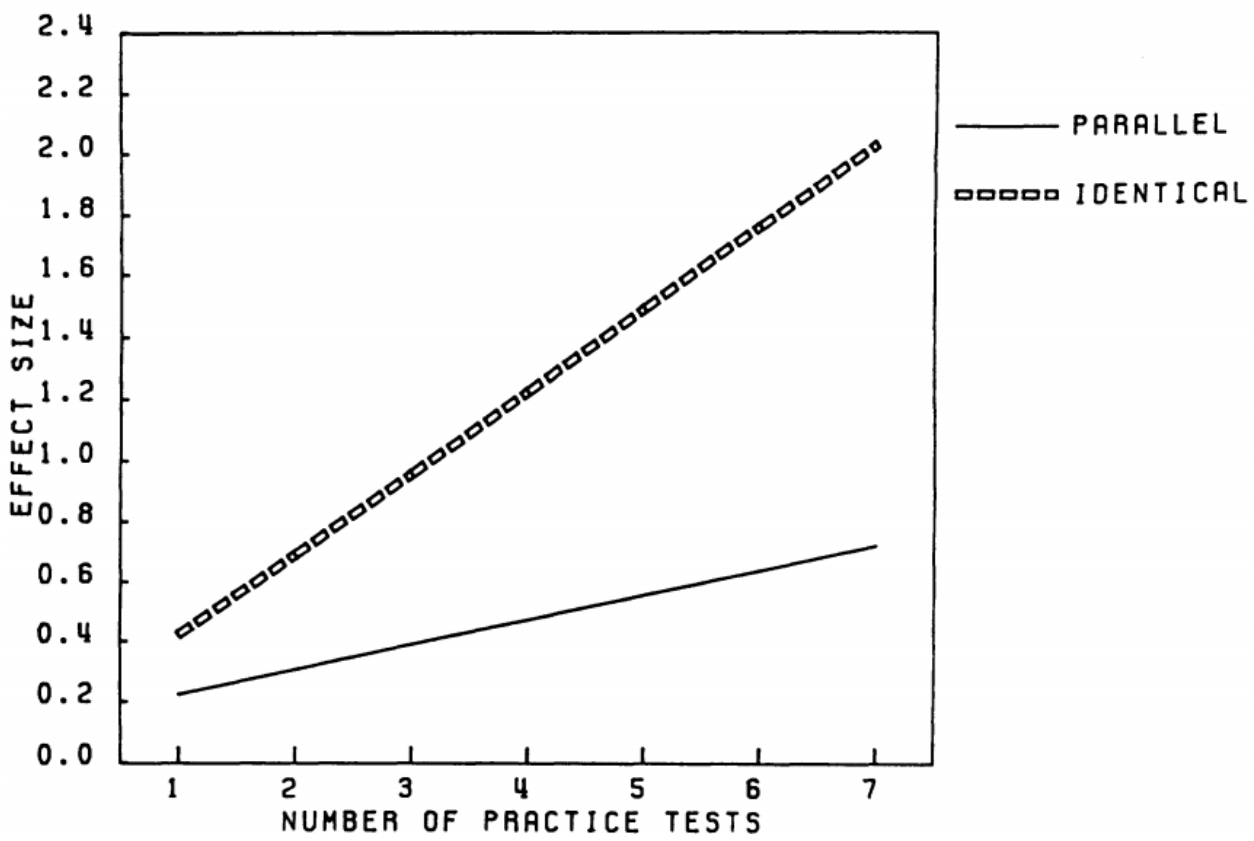


Figure 10.1: Effect sizes as number of practice exams increase, from Kulik et al [163], whether exams were identical or parallel to practice.

10.2 Methods

Retake opportunities were added to Physics 211, which is the target course of the preparation in Physics 100. The retakes were added in Spring 2018, the semester immediately following Fall 2017, when bi-weekly quizzes and retakes were added to Physics 100. Thus, students who took Physics 100 with frequent quizzes in 2017 would be the recipients of the change in Spring of 2018, along with their classmates who entered Physics 211 directly. The retakes remained part of the course through Spring 2019.

Physics 211 follows a similar content cycle to Physics 100; students begin with prelectures and checkpoints, attend lecture, then complete online homework and attend discussion sections. The online homework in Physics 211 is mainly immediate feedback format (as opposed to mastery), and discussion sections include an open response paper and pencil quiz each week. Rather than a single midterm and final, Physics 211 has three hour exams and a final exam. The course enrolls over 1000 students in the spring semester, which is historically about 20% Physics 100 alumni.

The hour exams are multiple choice via scantron, which are proctored in person with paper and pencil. There are usually around 25 questions to each exam, corresponding to 90 minutes. The exam retakes were proctored through the Computer Based Testing Facility (CBTF), the same as the bi-weekly quizzes. If a student decided to use their retake opportunity, they scheduled a time online through the CBTF and arrived at their scheduled time. The retakes were comparable in length to the hour exams and students were also given 90 minutes. As mentioned in the last chapter, tests at the CBTF display correctness information for each problem as students submit answers to complete their retakes.

Learning from our implementation of bi-weekly quiz retakes, both the exam and retake were weighted equally to ensure students were attempting both the original and retake exam seriously. If using the retake, students received the average of their two scores. Only students who scored below 60% on the original exam were encouraged to use the retake.

Each exam had corresponding practice exams, also informed by the demand for them during our bi-weekly quizzes. In 2018, the practice exams were optional exercises with one version which had supplied "help" videos alongside the problems; the videos were worked out narrated animated solutions, similar to the solution videos provided in mastery homework. There were three versions of the exam delivered through the online homework system, but other versions which could be printed and worked out by hand. We were able to collect data for students' interactions with the online practice exams, but did not have information about how they engaged with printed copies. In 2019, the format of the online practice exams were changed; help videos were not provided until students submitted answers and students were only allowed a single submission for their answer. This was done to make the practice exams more similar to an actual exam.

10.3 Results

10.3.1 Exam Retake Behavior and Performance

In the first year of exam retakes, Spring 2018, we saw that students self-selected to use retakes as we hoped; Figure 10.2 shows the fraction of retakes used, binned into 10 groups based on their exam scores. Around

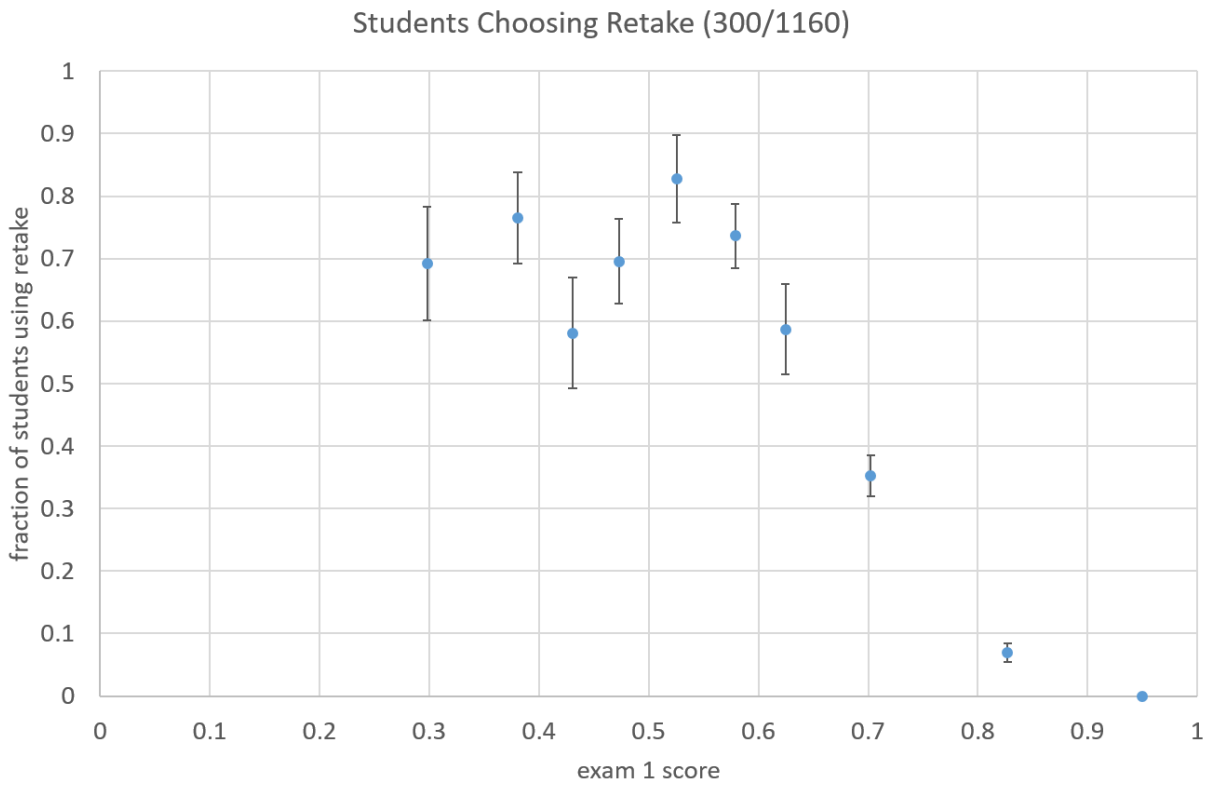


Figure 10.2: Fraction of students using retakes on exam 1, binned by exam 1 score.

70% of students who failed the exam opted to take the retake and the fraction of students using retakes declined as their exam scores improved. This is similar to the distribution we saw in the quiz retakes (Figure 9.6) but the cutoff from high retake use to low retake rates was earlier and more steep for Physics 211. This suggests that students were more conscientious about using retakes only if their score was low, possibly due to the risk of losing points on the retake, which was not a possibility in Physics 100.

On those retakes, students were able to bring their exam scores up about 10% on average. The distribution of gains is given in Figure 10.3, where we see the students peaking around the 10% improvement, with the majority in the -5% to 20% gains range.

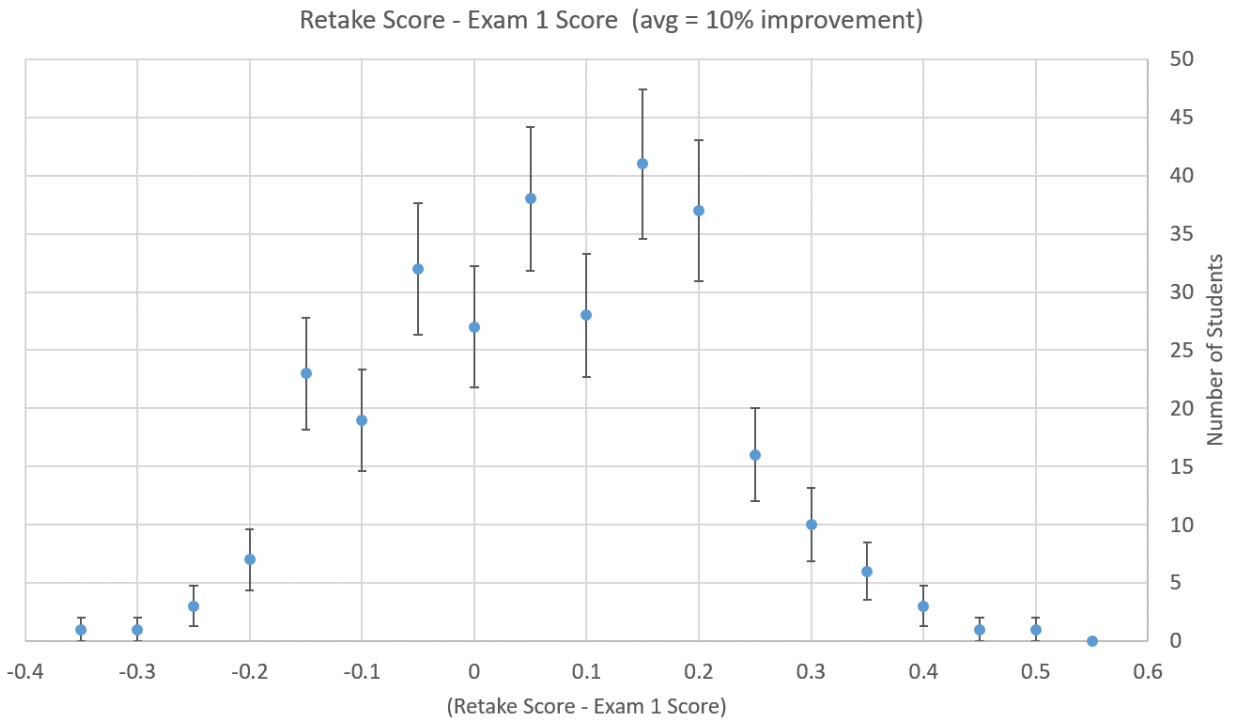
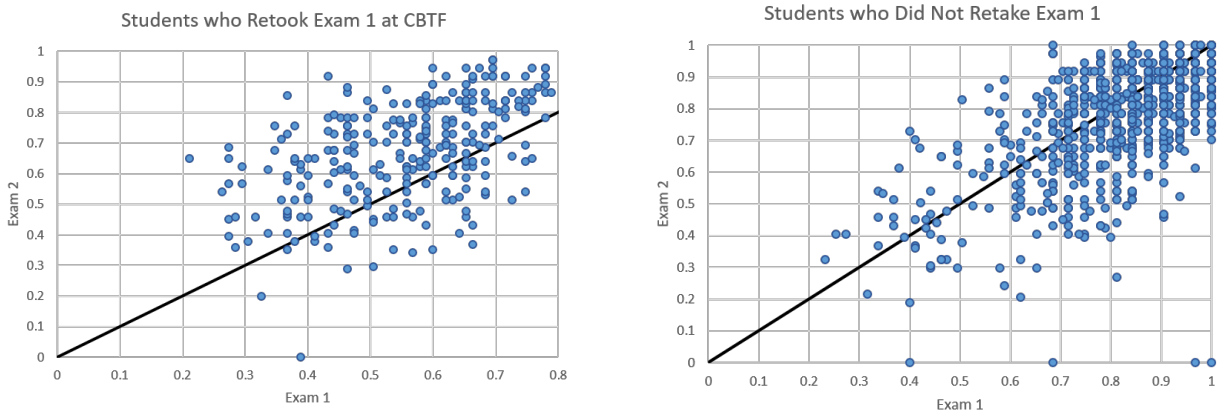


Figure 10.3: Binned distribution of student gains on retake over original exam score for students using retake on Exam 1.

10.3.2 Effect of Retakes on Next Exam

One hopes that the retake will be motivating for students to change their study habits between the first and second exam. To compare students who did and did not use retakes, scatter plots of students' performance on Exam 2 compared to Exam 1 are shown in Figure 10.4. On average, Exam 1 had a score of 76 $\pm$ 1% and Exam 2 had a score of 75 $\pm$ 1%, but students who used the retake exams scored 11% higher on Exam 2 (an average of 57% to 68%).

Although this is an encouraging result, improvements may be caused by self-selection; those who did not do well and were motivated to improve may have been more likely to opt in to a retake. We compared to a previous year, when retakes were not offered, and saw that the 300 lowest scoring students on Exam 1 also saw about a 10% increase from Exam 1 to Exam 2; the scatter plot for these students in 2017 is shown in Figure 10.5. The cutoff score for the 300 lowest scoring students in 2017 was a 64%. Students using the retakes in 2018 were not necessarily the lowest performing students, but 262 of the 300 students taking the retakes scored below 64%, making this a relatively good comparison. Looking only at our students who scored under 64% on Exam 1, there is still a 9% overall performance boost during the year with retakes. Those using the retake gained 10% on Exam 2, whereas students who did not use retakes scored 6% higher than their score on Exam 1.



(a) (b)

Figure 10.4: Scatter plots of Exam 2 scores compared to Exam 1 scores for (a) students who used the retake exam for Exam 1 and (b) students who did not retake Exam 1.

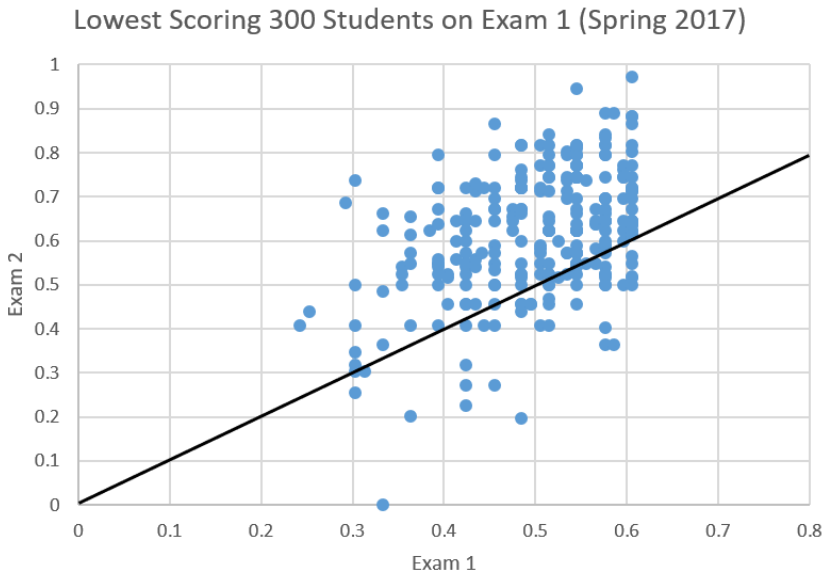


Figure 10.5: Scatter plot of performance of Exam 2 by performance on Exam 1 for the lowest scoring 300 students in Physics 211 during Spring 2017, when retakes were not offered.

The effect of retakes in general on students' final grades was evaluated by checking the course grades of the lowest scoring students on Exam 1 in both Spring 2017 (no retakes) and Spring 2018 (retakes offered). The distribution of these students' final course grades are shown in Figure 10.6. We saw that the retakes had no significant effect on students' long-term performance in the course.

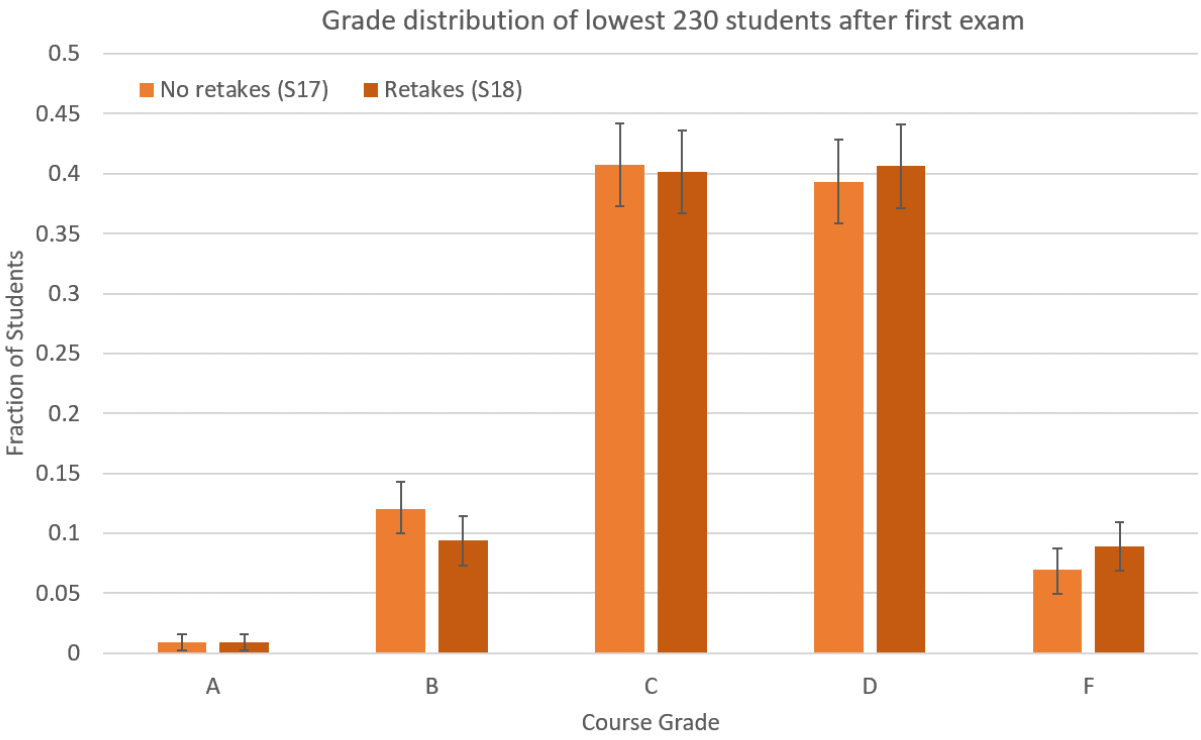


Figure 10.6: Final course grade distribution for the lowest scoring students on Exam 1 in Physics 211 during the semester without and with retakes offered.

10.3.3 Use and Effect of Practice Exams

In the same vein as retakes, we were interested to see how the practice exams were affecting students' exam scores. The practice exams were delivered to students differently in 2018 and 2019, the change inspired by our results in 2018.

In 2018, when the practice exams were given as optional immediate feedback problems (and one version supplied with narrated animated solutions as help, alongside the problems), we saw that students' engagement with the online practice exams had no correlation to their performance on the actual exam. There were three versions of practice exams online, corresponding to about 80 questions combined. Figure 10.7 shows the distribution of students' scores on the exam based on the number of practice problems they completed. Even students finishing nearly every online practice exam still had exam scores ranging down below 40%.

On the practice exams in 2018, students' performance was not a good predictor of their actual exam perfor-

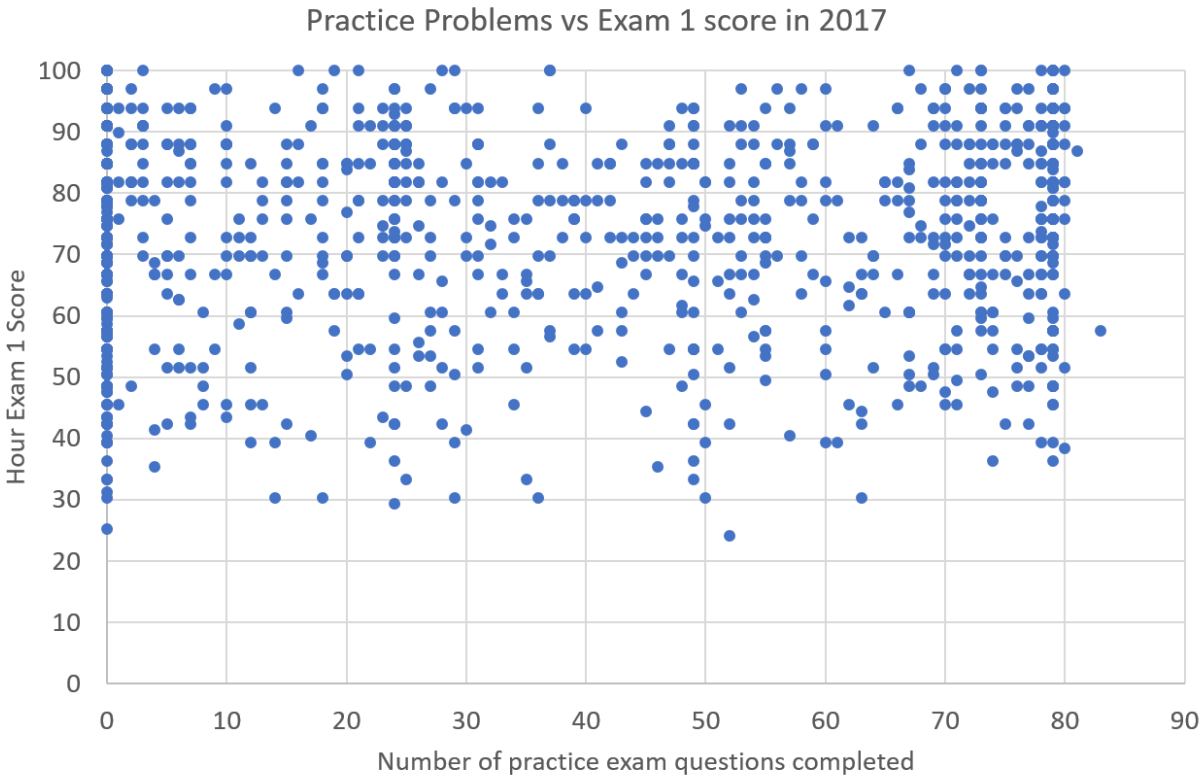


Figure 10.7: A scatterplot of students' hour exam 1 scores compared to the number of practice exam questions they completed.

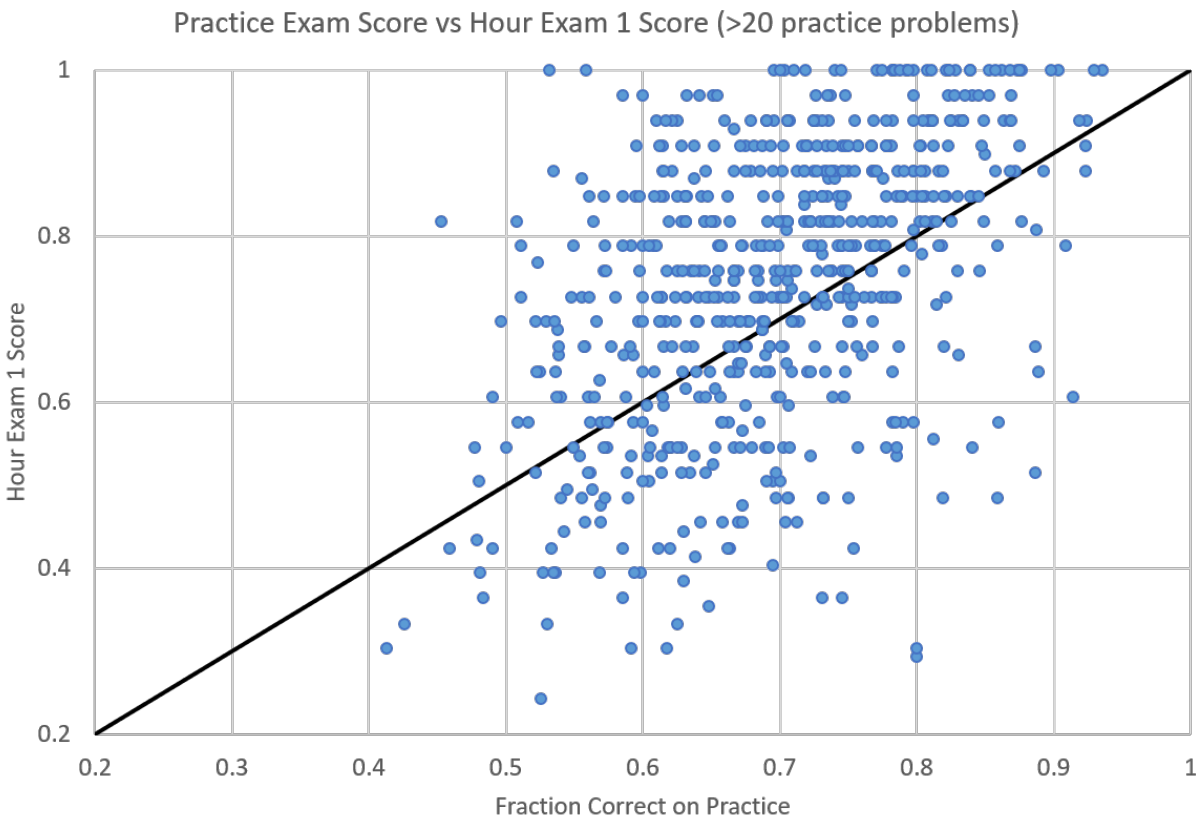


Figure 10.8: Students' hour exam 1 scores compared to their fraction of correct practice problems in 2018, for students who used at least 20 practice problems.

mance. Figure 10.8 shows students' exam scores compared to their practice exam scores for students who attempted at least 20 practice problems (nearly one practice exam). About a third of students performed worse on the actual exam than the practice. From looking at students' logs from the practice exams, many students submitted multiple answers within seconds of each other, suggesting that they were not using the practice exams conscientiously.

The delivery method of the practice exams was changed for Spring 2019 in Physics 211; rather than supplying solutions beside the problems, they were given to students after submission, and they were only allowed a single submission for each problem. By changing the delivery format of the practice exams, they became much better predictors of students' final exam scores, shown in Figure 10.9. Additionally, students improved over the course of their study with the practice exams. Recalling that a single practice exam is around 25 questions, Figure 10.10 shows students' scores on their first 24 questions (the first practice exam) compared to their score on the rest of their practice exam problems. Students who attempted at least two practice exams saw an improvement of 8% from their first practice exam to the rest of their practice.

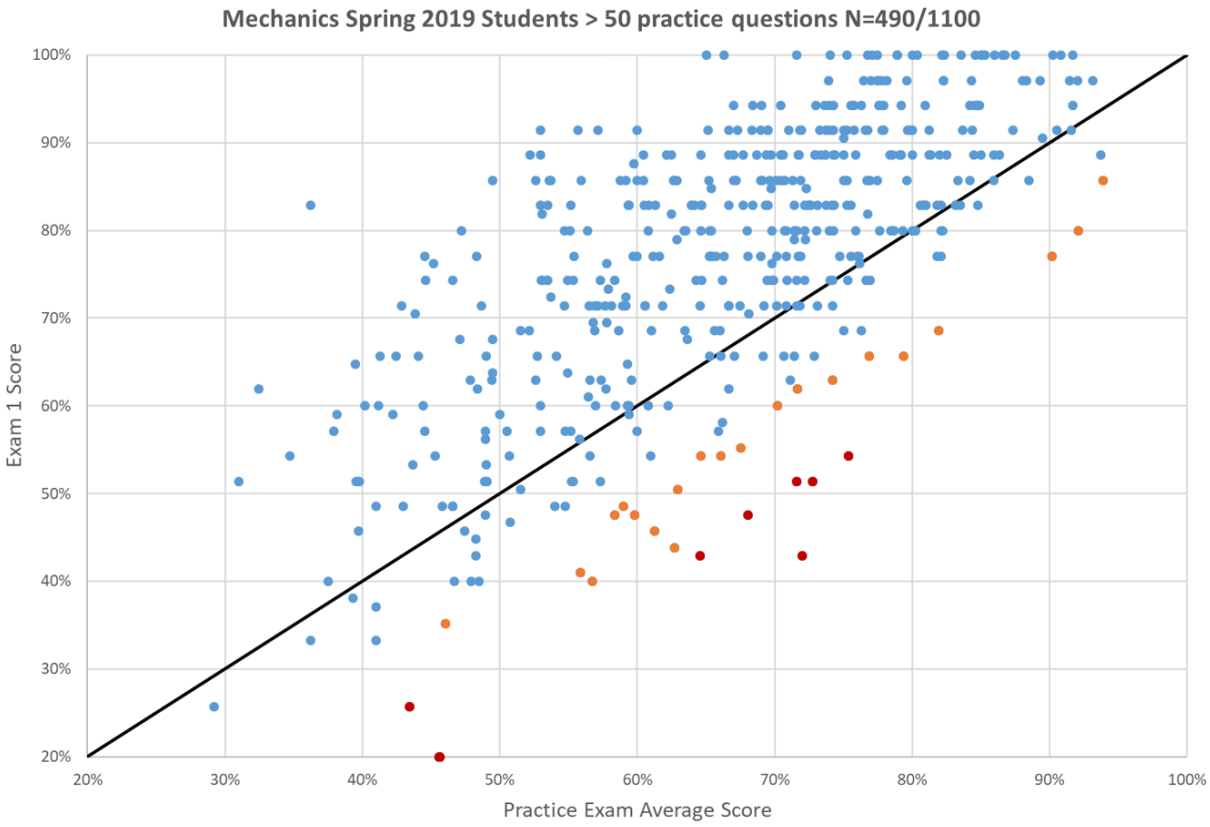


Figure 10.9: Students' hour exam 1 scores compared to their fraction of correct practice problems in 2019, for students who used at least 50 practice problems. Orange points correspond to students scoring 2 standard deviations lower on the exam than their practice. Red points correspond to scoring 3 standard deviations lower.

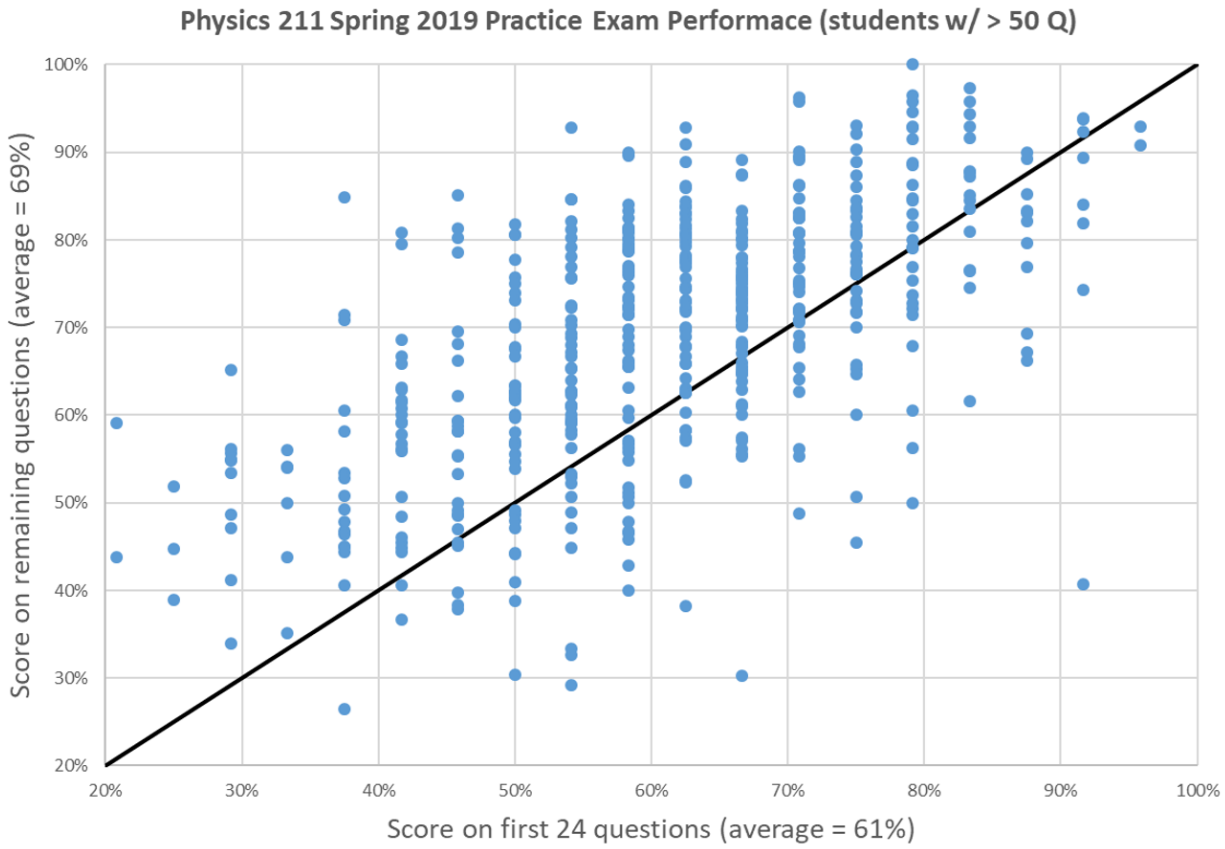


Figure 10.10: Scatter plot of student scores on their first practice exam (24 questions) compared to later practice exams, for students in 2019 who submitted answers for at least 50 questions.

10.4 Discussion

Both retake exams and practice exams can be components of frequent testing, but we saw that practice exams had a larger impact on student performance. On the retake exams for students in Physics 211, we saw that increasing the stakes of the retakes from Physics 100 (no risk on quiz retakes) to Physics 211 (exam retakes can reduce score) affected the fraction of students using retakes, especially reducing the number of high performing students using the retakes. The retakes in both courses were intended for students who scored poorly on their first test, so encouraging students to take retakes intentionally rather than by default was a good result. One also hopes that making an intentional decision to use the retake and the additional stakes would encourage students' study between the original and retake exams; students scored an average of 10% higher on their retake exams compared to 7% in Physics 100 quiz retakes. As in Physics 100, retakes helped lower scoring students approach their classmates' scores, but it is unclear whether these improvements are primarily corrections from statistical fluctuations via a regression toward the mean; it is possible that the result is largely from students who performed below their ability on the first exam and were given an opportunity to show their expertise on the retake.

From their first exam to their second exam in Physics 211, students who used retakes scored 11% higher despite the average exam scores for Exam 1 and Exam 2 being comparable. However, students using retake exams tended to be lower-performing students and we saw that the lowest performing students in 2017 (when there were no retakes) also saw about a 10% increase from Exam 1 to Exam 2. This could be attributed to the frequent testing model; students who had already done poorly on one exam could be more likely to change their behavior for the next exam, and self-selection could be driving the result. Alternately, this improvement could also be a statistical regression toward the mean. The lowest performing students on the first exam saw no significant difference on their final course grades in the years with and without the retake exams, so we believe that the retakes were successful at smoothing out fluctuations but not necessarily successful at improving student learning.

The practice exams, on the other hand, showed student learning improvement and better tracked with students' exam grades, when they were delivered in a format to simulate actual exams. In 2018, when the practice exams were given alongside help videos and allowed students to submit multiple answers, students' participation and performance on the practice exams were not correlated with their actual exam scores. Giving multiple attempts at problems may be sending an incorrect message; students who are able to solve problems with help videos and multiple attempts may falsely think they are able to recreate the problems on the exam in a single attempt with no notes or help. This could be explained by "the illusion of understanding," where students reading or watching worked examples interpret familiarity as competence, but cannot recreate problem solutions that they believe they understand [65, 165–167]. Schroeder saw that students who self-reported full confidence in understanding a worked example still were unable to successfully complete a the exact same problem within the same class session [65]. Chi et al cite that the illusion of understanding can be tempered if students who are asked to actively use an explanation to do another task [168], but the multiple choice and multiple attempt nature of the practice exams did not require students to demonstrate their competence after using help videos.

When practice exams were delivered in 2019, forcing students to only submit one answer and withholding solutions until they had made an attempt may have given students a better sense of their competency with the material and required them to engage more actively with the problems. Their performance on these

practice exams was correlated better with their actual exam scores, and students saw improvement from their first practice exam to their next. The practice exams were still available alongside notes and had no time limit, making them more lenient than actual exams, but a better formative assessment than in the previous year.

When considering frequent testing, the results from our retakes and practice exams in Physics 211 suggest that practice exams, when implemented as close to testing conditions as possible, were better for student learning. In particular, grading the practice exams more like an exam gave valuable formative assessment to students. Rather than inflating their sense of competency, such as in 2018 when students performed much better on the practice exams than the actual exams, in 2019 students were able to see improvement in their practice exams and have that movement reflected in their exam scores. In Physics 100, students requested pretests for their bi-weekly quizzes, and this study provides evidence that their addition is valuable.

Chapter 11

Values Affirmation Replication

11.1 Introduction

Over the years of implementing tools for students in Physics 100, a major lesson learned was the importance of students' attitudes and affect when engaging with material. Students' attitudes about physics and sense of belonging in the field are influenced by their identities and experiences around their identities within the field and beyond, particularly for students from marginalized populations [169–182]. These differences between majority and minority populations in physics can manifest in students' sense of self-efficacy and physics identity, which tend to persist from elementary school [178, 183] to high school [183, 184], introductory courses [169, 170, 175, 178, 185], upper-level courses [170], graduate school [170] and eventual careers in physics. In particular, students' sense of physics identity is a strong predictor of their persistence in the field [179, 181, 186–191]. A major focus of the projects described in this dissertation is improving retention of physics and engineering students, so attention to factors beyond course content and structures is extremely relevant.

This chapter documents a replication study of a values affirmation exercise that received considerable attention for reducing performance differences between men and women in an introductory physics course at the University of Colorado, Boulder [192]. Values affirmation exercises typically consist of a short writing activity about self-identified values, which is meant to protect minoritized students against threats to their identity and in turn bolster their self-efficacy and performance.

11.1.1 "Achievement Gaps" in Physics

Many studies which document student differences between majority and minority populations within academic fields cite performance in the form of achievement gaps with respect to gender and race [175, 176, 185, 193]. This method of comparing groups' performance has been criticized or cited with caution by some education researchers [44, 172, 176, 186, 194, 195]. Guiérrez in particular cautions against studies which only document the existence of the gaps, asserting that these studies provide a static picture and do little to provide insight into addressing the complex problem [194]. Additionally, comparing majority and

minority population students inherently reinforces a deficit model; differences between groups are framed as inadequacies in the minority populations [172, 176, 194] and assume that acting like the majority group is the goal. For example, Traxler et al explain that gender gap analyses often explicitly or implicitly ask the question, "why can't women be more like men?" [176]. Often comparisons are made by performance on standardized measures which have been validated on physics student populations which are primarily white men; there has been little effort to validate these measures with attention to marginalized students [176]. In response to some of these criticisms, Gloria Ladson-Billings offers a re-framing of the "achievement gap" as an "education debt" to emphasize that differences between students' performance and self-efficicacy are not faults of the minoritized students, but rather an accumulation of historical, economic, sociopolitical effects [44].

Furthermore, achievement gap research tends to erase the unique experiences of those at intersections of identities by treating all women or all students of color, for example, as homogeneous groups [176, 179, 194]. The theory of intersectionality, coined by Kimberlé Crenshaw, cites the damage of considering the consolidation of people along a single axis of identity by hilighting instances where the practice has hurt Black women [196]. She cites examples where initiatives and policies intended to help women ended up only serving white women, and policies intended to help Black people served to only help Black men [196]. In physics education, the majority of research is conducted at large primarily white institutions within mostly male classes, and thus we are at risk of also oversimplifying and prescribing interventions which overlook more marginalized people. Similarly, when considering gender gaps, researchers often treat gender as a purely binary variable [176], based on admissions information, which erases the experiences of gender non-conforming students and misrepresents transgender students who may be on record with the university as their wrong gender.

Alternately, Lubienski stresses that it is "irresponsible" not to investigate achievement gaps [197], and Traxler et al cite that some thoughtfully conducted gap analyses have helped the education community understand mechanisms which contribute to performance differences [172]. Guiérrez advises that it is more useful to focus on improving learning within marginalized populations without needing to compare them to majority students, and whatever benefit gap analyses bring come at a cost. Thus, it is important to be mindful when citing and studying performance gaps.

11.1.2 Stereotype and Identity Threat

Of the dangers of "gap" literature, one is its oversimplification of many entangled contributions to individuals' experiences and performance. However, education research has attempted to isolate variables that affect students' performance and senses of efficacy within classrooms; there is strong research which documents the effects of identity threat and stereotype threat, over 300 studies in peer reviewed journals [198–204].

The effects of stereotype threat were named and introduced by Steele and Aronson in 1995 [205] in studies that saw Black middle school students scored below their ability on tasks when they perceived a threat to reinforce negative stereotypes. The studies saw that all participants scored similarly when the task was framed as a non-diagnostic study of problem solving, but the group which had the task framed as a diagnostic of intellectual ability saw differences in performance between White and Black students, which the authors attribute to the threat on Black participants to be concerned about stereotypes of intellect and reinforce

those negative stereotypes [205]. In one of the studies, they asked participants to fill in letters to complete words; Black students in the group that had the task framed as a diagnostic intellectual task were more likely to complete the words into ideas related to stereotypes and self-doubt. For example, the prompt L A _ _ could be completed to LAZY, W E L _ _ _ _ to WELFARE, F L _ _ _ to FLUNK [205]. The authors saw that these words appeared drastically more often for Black participants in the stereotype threat condition, but not in the non-diagnostic condition, which they use to justify their assertion that the threat was preoccupying cognitive resources [205].

These results have been replicated along many axes of stereotype in academic settings, including women on math tasks [206–210], students from low socioeconomic backgrounds [211, 212], and Hispanic or Latinx students [213–215]. The effects are not confined to academia, and have also affected memory tasks in the elderly [216], anxiety in gay men providing childcare [217], White men in sports [218], and women in negotiations [219] and driving [220]. In two meta-analyses which combine data from nearly 19,000 students, Walton and Spencer quantify the degree of underperformance by stereotyped students to conservatively be near 0.2 standard deviations, which they assert contributes a large part of the "gaps" seen in literature [199].

The mechanisms which drive stereotype threat are often cited as being cognitive. Increased anxiety and distraction by the threat can use up cognitive resources which would otherwise be focused on the performance tasks [199, 200, 205, 221, 222]. In particular, Krendl et al used MRI scans to verify that women doing math in a stereotype condition had less activation in their brains in areas related to problem solving than women in the non-threat condition, and saw higher brain activity in a region associated with processing negative information [223]. Likewise, Croizet et al measured heart rate fluctuations associated with mental workload and also discovered students facing stereotype threat were experiencing higher mental workload which taxes students' limited working memory [224]. In a synthesis of research of stereotype threat related psychology, Steele et al expand this explanation by citing the cyclic nature of stereotype threat; students who perform poorly are treated differently by teachers, adaptively adjust their self-image to reflect the failures, and different opportunities are available to them which reinforce the stereotype and threat [200]. This can affect motivation and students' performance expectations [225–227], which in turn influences students' behavior.

Steele et al cite a number of factors and strategies which can temper the effects of stereotype threat [221]. Many of the strategies are relational; students are better situated to buffer against stereotype threat if they have support systems to directly or indirectly counter stereotypes. Steele et al argue that having friends within their identity group can reduce students' endorsement or anxiety around stereotypes simply by having more interactions with people which can contradict stereotypes and that the endorsement of stereotypes influences how students are affected by stereotype threat [221]. Graham et al speak particularly to the threat to Black students in Predominantly White Institutions (PWIs); they found that students who had friend groups which were made of primarily Black students in high school showed more ease in transitioning academically and socially into college (particularly into a PWI), even controlling for socioeconomic status and neighborhood [228]. Likewise, seeing successful people of their identity can reduce the threat of stereotype by providing counterexamples of success [221]. Mentors or teachers, regardless of the alignment of their identity to students can also reduce threat by valuing students' contributions and especially by speaking explicitly about stereotypes [221]. Even the act of explicitly citing an assessment as being neutral to stereotypes was shown to reduce stereotype threat for women on math tasks [206]. Within students themselves, the belief of

malleability of intelligence such as Dweck's growth mindset, can reduce the impact of stereotype threat by rejecting "innate" qualities of intelligence associated with stereotypes [221, 229, 230].

11.1.3 Self and Values Affirmation History and Theory

With the adverse affects of stereotype threat explicitly documented, many education researchers have prioritized energy towards addressing the threat. These efforts include social psychological interventions which target students' thoughts, feelings, and beliefs about themselves. Values affirmation exercises have received considerable attention, in part because the exercises have shown long lasting effects at minimizing "gaps" with only brief interventions. These interventions usually consist of short writing exercises where students identify values that are important to them and reflect on their personal importance.

Values affirmation exercises are a type of self-affirmation, a method described by Steele in which people respond to a threat to their integrity by fortifying another aspect of the self, which does not necessarily need to be related to the threatened domain [231]. He cites a study of housewives asked to help on a community project: some were told that it was known within the community that they were not generally "cooperative" in community projects, and these women were more likely to volunteer in a followup by a completely different person in a separate encounter. When their self-perception of themselves as helpful people was threatened, the women were more likely to seek out another venue to restore their sense of self and integrity [232]. In the same vein, students who are feeling threatened in their academic environment can bolster their sense of integrity by validating aspects of themselves that are valuable to them [233]. In their review of self-affirmation literature, Sherman and Cohen speak to this with direct application an educational environment [234]:

> When global perceptions of self-integrity are affirmed, otherwise threatening events or information lose their self-threatening capacity because the individual can view them within a broader, larger view of the self. People can thus focus not on the implications for self-integrity of a given threat or stressor, but on its informational value. When self-affirmed, individuals feel as though the task of proving their worth, both to themselves and to others, is "settled." As a consequence, they can focus on other salient demands in the situation beyond ego protection.

Values affirmation exercises do not necessarily alleviate specific threats, but instead reduce the power of the threat to a person's adequacy [200, 202, 213], freeing up cognitive load to focus on the tasks at hand. Shnabel et al examined different types of emphases in values affirmation exercises and discovered the strongest effects in students who were prompted to write specifically about social belonging. They prompted these reflections by asking students to write about experiences which made them "feel closer and more connected with people" [203] which aligns with Steele et al's emphasis on relational counters to stereotype threat [221].

Some of the earliest values affirmation experiments were conducted by Cohen et al [235], who saw a reduction by 40% in the achievement gap for Black students compared to their White classmates. In their followup study, they tracked the same students over two years and saw that the effects resulted in higher GPA (0.24 grade points higher than students who did not receive the affirmation intervention) and a lower rate of remediation or grade repetition (5% compared to 18%) [236]. In the same way that stereotype threat can compound, they stress that values affirmations have the potential to persist recursively; students who feel valued may perform better, which affects how they see themselves and how they are treated, which can

amplify the effect with each new validating event [236]. Values affirmation interventions are intended to disrupt the cycle of stereotype threat, either slowing its effects or setting in motion a positive counter cycle [200, 202, 236].

Because of their recursive nature, the benefits of values affirmation exercises are affected by their timing; earlier interventions can disrupt and bolster students against recurring interactions which may threaten their self-worth [200]. Sherman et al suggest that administering interventions during times of transition are especially effective (such as students entering middle school, high school, or college) because the nature of these transitions often include increased academic expectations, flux of identity, and disrupted support systems for students [213].

Additionally, Yeager and Walton notice that values affirmation exercises are most effective when students are unaware of their true intention [202]. By labelling interventions as "interventions," students are receiving a stigmatizing message that they are in need of help (which actually plays into stereotype threat). Instead, students who feel control over the activity are more able to take ownership of their success rather than attribute it to a "heavy-handed intervention," amplifying the recursive nature of growing self-efficacy [202]. Sherman et al gave the additional rationale for subtlety in delivery: students who are aware that the intervention is meant to bolster self-worth will be using cognitive resources to consider the benefits of the intervention as a means to an end, rather than fully engaging in the activity and focusing inwards [237]. Their study explicitly measuring how awareness affects values affirmations showed empirically that awareness attenuates the positive effects of values affirmations [237].

11.1.4 Values Affirmation Experiments and Replication

As mentioned previously, the earliest values affirmation experiments were conducted by Cohen et al [235, 236], which showed promising results in elevating Black middle school students' performance. Since then, others have successfully implemented values affirmations exercises in their labs or classrooms with success. Sherman et al found that a values affirmation intervention improved grades for Latinx American students who participated in the writing activity [213]. Another study intending to replicate Cohen et al's studies was done in a predominantly Black and Latinx middle school, with 80% of the students enrolled in a free lunch program. Values affirmation exercises were distributed to half the students through 24 homeroom teachers, and results showed that students who participated in the affirmation exercise outperformed students in the control condition. This study did not include a control for White, middle class students because the sample was too small, but instead focused on the ability of the affirmation to lift performance of stereotyped students [238] as suggested by Guiérrez [194]. Values affirmation exercises were also successful for ethnic minority medical students [239], first generation students in biology [240], LGBTQ students (in conjunction with another activity) [241], and women in science [192, 242].

Alongside successful replications, the literature also includes studies with mixed or null results [243]. When delivered to students in three St. Paul middle schools, values affirmation exercises did not statistically affect students' performance on a Minnesota state standardized test, except for women on their math scores. The researcher expected to see an improvement for racially minoritized students but was surprised to see no effect [244]. Similarly, when the intervention was given to students in six middle schools in the Philadelphia area, the author cites that the replication failed to have significant results, and actually hurt the performance of

female students [245]. Another study across eleven middle schools saw some improvement, but cited that the results were much smaller than predicted by Cohen et al's results [246]. In the college-level domain, an affirmation study in introductory biology courses marginally helped racial minority students but the researchers also saw that the intervention negatively affected White women, which contradicts other studies [247]. In physics, biology, and biochemistry courses, a replication of a successful gender oriented values affirmation saw no difference between treatment and control conditions for men or women, though they attribute the null result to limited initial differences between men and women at their institution [248].

Most of these published works showed some mixed results, and it is possible that replications which saw purely null results were not published. Franco et al followed 221 funded social science studies and discovered that the majority of null results are not even written up (over 60%), and those that are written are less likely to be accepted for publication than stronger results [249, 250]. This publication bias has been persistently increasing across fields and across different countries [251], especially in the social sciences and biomedical research [251–253]. The exclusion of null results skews literature by selectively amplifying positive results and dismissing null results which contradict them [250, 253, 254]. In the extreme case when one is using p-values of 0.05 as significance, if 5% of "chance" positive results are published and 95% of null results are not, our understanding is extremely skewed [253, 254]. In fact, a synthesis of studies by Sterling et al saw that over 95% of studies biomedical and social science journal articles cited statistically significant results [255]; either the proportion of significant studies has increased or bias against non-significant results in publishing is driving these numbers up [254]. In the case of values affirmation interventions, the experiments are relatively quick, allowing for easy repetition and replication; statistically, more research groups attempting to replicate a result will increase the chances of statistical fluctuations that can be published as positive results.

Within physics education, a notable positive result by Miyake et al at the University of Colorado, Boulder showed statistically significant results in reducing the gender gap in introductory physics, both in grades and scores on the Force and Motion Conceptual Evaluation (FMCE) [192]. The result was published in 2010 in *Science* and has been cited over 500 times at the time of this dissertation. A subset of the authors of the original paper failed to replicate these findings; they saw reduction in differences between men and women on exams but the interaction of the treatment and gender was not statistically significant and the results of the FMCE showed women performing worse with the affirmation treatment [256]. In contrast to the publicity of the first study, this contradictory replication has been cited on the order of 20 times. The original study is often cited in physics and STEM education literature without qualification or as a strong result to inform teaching strategies [174, 257–265]. Some researchers acknowledge the mixed results [176, 193, 201, 266, 267] or caution about their context-dependency [243, 248], but simply by virtue of numbers, the first study is often cited without reference to the less successful replication. The importance of subtle contexts which affect the implementation of values affirmation studies are asserted by Bradley et al [243], and Conlin et al stress that the absence of attention to null results deprives the research community of opportunities to explore how these contexts can be refined [254]. As a large Research 1 university with high-enrollment physics classes (and thus, statistical power), the University of Illinois at Urbana-Champaign is well-positioned to either replicate or define boundaries of values affirmation interventions.

Table 11.1: Activities associated with Values Affirmation Replication: timing, context of implementation, and notes for clarification.

Timing	Activity	Context	Notes
Week 0	Professor primes students to expect activity	Lecture	Professor emphasizes the importance of science communication and tells students to expect writing exercise in discussion.
Week 1	First Values Affirmation writing activity	Discussion section	Activity and TA scripts identical to CO implementation. Consent form adapted from CO implementation.
Week 2	Attitudes survey to test stereotype endorsement	Online Checkpoint	Questions to test stereotype endorsement embedded in the CLASS. Full CLASS and 3 questions to test endorsement for Physics 100. A shortened version of the CLASS and 2/3 of endorsement questions for Physics 212.
Week prior to Exam 1	Second Values Affirmation writing activity	Online homework	Activity identical to CO implementation, delivered online. Week 10 for Physics 100. Week 4 for Physics 212.
Exam Week	First Exam	Proctored exam	Week 11 for Physics 100. Week 5 for Physics 212.

11.2 Methods

Our values affirmation replication was done in two introductory physics courses at the University of Illinois, Urbana-Champaign. The intervention was implemented in Physics 100, already discussed extensively, and Physics 212, a calculus-based introductory electromagnetism course. Physics 212 is typically taken by engineering students in their second year. These students have already successfully completed Physics 211, the introductory mechanics course, meaning they are overall likely stronger students as they have the experience of completing at least one other physics course and are selectively the students who were able to succeed in that course (As we saw earlier, there is some attrition of students through the introductory physics courses.). Of course there is variation of experience within all course populations, but we expected the two courses to be enlightening populations for checking the effects of values affirmation exercises. Both instructors for Physics 100 and Physics 212 were men.

The timing and implementation of the values affirmation exercises at the University of Illinois mirrored the implementation of the exercises at the University of Colorado. This author was in contact with Professor Miyake, the first author of the Colorado affirmation paper, and received an extensive implementation guide for replication which was followed as closely as possible. The timing and activities are outlined in Table 11.1, which will be elaborated below.

11.2.1 Writing Activity

To maintain the "stealth" aspect that has historically improved the effectiveness of values affirmation exercises [237], the values affirmation exercises were emphasized as being "writing" exercises. The exercise is first introduced to in students' first lecture and stressed as being important for science communication, as recommended by the implementation guide.

The writing exercise itself was given to students in their first discussion (recitation) section by their Teaching Assistants (TAs). Prior to the first week, the researcher met with the TAs of the two courses to distribute

the exercises, which were paper and pencil packets in plain envelopes, one for each student. The TAs were given the number of exercises needed for their sections which included a random sampling of Control and Treatment conditions that could be handed out randomly, blind of condition. TAs were not told the nature of the experiment, but were given a script with answers to potential student questions; most importantly, they were told to ensure students knew that the activity was not graded (to reduce stress that could compromise the activity) and that people associated with the course (themselves and the professor) would not read their responses. The teaching assistants were not told the true nature of the experiment to reduce the chance that their distribution and responses to students would unintentionally give students clues to its purpose. This was suggested by the implementation guide from Colorado and the script was identical to the one used in their experiment. The decision to have TAs distribute the exercise was also suggested, as it reduces emphasis that the activity is an experiment and allows students to focus on the activity. As required by our Institutional Review Board, our consent forms attached to the activity were slightly more specific about the research aspect than the provided template, but also did not hint at the goal of the activity to measure or influence gender differences.

Students were given 15 minutes to complete either a control or treatment writing exercise, which had three parts:

Part One Students were given a list of personal values and asked to circle two or three that are **most** (treatment condition) or **least** (control condition) valuable to them. The options listed were identical to the Colorado implementation [192] and Cohen et al's original implementation [235]. The listed values were:

- Being good at art
- Learning and gaining knowledge
- Relationships with family and friends
- Government or politics
- Independence
- Creativity
- Athletic ability
- Music
- Belonging to a social group (such as your community, racial group, or school club)
- Spiritual or religious values
- Sense of humor
- Career

Part Two Students in the treatment condition were asked to think about the values and a time that the values were or would be important to them. Students in the control condition were asked to think about why the values might be important to someone else. All students were encouraged to focus on their thoughts, rather than grammar, spelling, or quality of writing, and write a few sentences about their importance.

Part Three Students were asked to look over their values chosen and reasons, then list each of their values again with the top two reasons the values are important to them (treatment) or the top two reasons they could be important to someone else (control).

The envelopes were collected by TAs and returned to the researcher. An outside researcher tabulated each student's condition (treatment or control) and consent given, which was used to generate a list of email

addresses to ensure that students would receive the same condition on the second iteration of the writing activity.

11.2.2 Stereotype Endorsement

About a week after the writing activity, students completed an online learning attitudes assessment with questions embedded to gauge their endorsement of the stereotype that men do better than women in physics. This needed to be done after the affirmations activity but as early as possible to avoid activating stereotype threat immediately before the exam. The learning attitudes assessment used was the Colorado Learning Attitudes about Science Survey (CLASS) [268], but the assessment was primarily a vehicle to distract from the questions about gender. All the items in the survey are given as agreement with statements on a 5-point Likert scale between "Strongly Disagree" and "Strongly Agree." The stereotype endorsement statements embedded into the CLASS are listed below:

1. According to my own personal beliefs, I expect men to generally do better in physics than women.
2. According to general beliefs in society, men are expected to be better at physics than women.
3. I think my physics teachers expect women to do better than men in physics.

All 42 questions on the CLASS plus the three additional endorsement statements were given to Physics 100 students, and a subset of the CLASS (18 statements) and two stereotype endorsement statements were given to Physics 212. The modified version was given to Physics 212 students because of length concerns by the professor of the course. The study by Mikake et al [192] found Question 1 (about personal beliefs) to be most useful for their analysis, so the first two endorsement questions were the ones included in the subset survey. Students received points for completing the learning attitudes survey but not for any specific answers.

11.2.3 Writing Activity Repetition and Exam

The second iteration of the writing activity was done online, the same as in the Colorado implementation. Two online forms were created with the same parts as in the paper and pencil implementation, one for control (asking students to reflect on values which are not important to them) and one for the treatment (asking students to reflect on their own values). The forms were sent to participants to match their condition in the first implementation; students received points for completing the activity, but not for any specific answers. Students who did not participate in the first activity (who enrolled late or missed the first discussion section) were randomly assigned one of the two conditions.

The timing of the second implementation was one week before each course's first exam. Since Physics 212 has three exams and a final, this came early, in Week 4, for Physics 212 students. Physics 100 has only a midterm before the final exam, and their second writing activity was in their homework in Week 10. The corresponding exams were given in Week 5 and Week 11, respectively.

11.2.4 Intended Analysis

The analysis done by Colorado included students' exam grades (immediately following the second iteration of the writing activity), students' FMCE scores, final grades in the course, and both exam grades and FMCE scores as a function of high or low endorsement of gender stereotypes in physics, as measured by embedded questions on the CLASS. In our analysis, we will repeat each of these measures except for FMCE scores, which were not included in our courses. The FMCE would only be relevant for content in Physics 100, not for Physics 212, so it was not included.

The rest of the analysis was repeated, and the exam scores as a whole and by gender stereotype endorsement were additionally refined using bootstrap analyses. The bootstrap analyses were done to better understand the statistical sensitivity of our results. To the extent that the student performance distributions are normal, the bootstrap analyses should be consistent with standard analysis techniques. Each bootstrap analysis calculated the size of the improved score (Treatment Condition - Control Condition) for each group of students by simulating 10,000 trials. From the boostrap, we calculated confidence intervals to give boundaries to the effect of the intervention.

11.3 Results

Tabulating responses after the writing activity, we had 389 participants from Physics 100 and 1012 participants from Physics 212, which are reflective of the different sizes of the course. Randomly distributing exercises effectively split participants into treatment and control conditions of comparable size; the breakdown of the number of students in each condition by class is given in Table 11.2. For comparison, the first Colorado implementation was done with 439 students and its second implementation included 363 students. In both these cases, these participants were a subset of the course (73% and 60%, respectively) [192, 256]. Our samples included only students who completed both exercises, so consist of a subset as well; 76% of Physics 100 and 92% of Physics 212 are included in analysis.

Table 11.2: Number of participants in each condition, by course, for Values Affirmation experiment.

	Physics 100 (N=389)		**Physics 212 (N=1012)**	
	Treatment	**Control**	**Treatment**	**Control**
M	121	120	366	381
F	75	73	143	122

11.3.1 Exam Results

Raw Data

The first result presented by the Miyake et al paper showed a significant reduction in gender differences for students in the treatment condition, which was presented as a bar chart and is shown for reference in Figure 11.1 [192]. Their scores used were adjusted to account for baseline ACT/SAT math scores, which increased

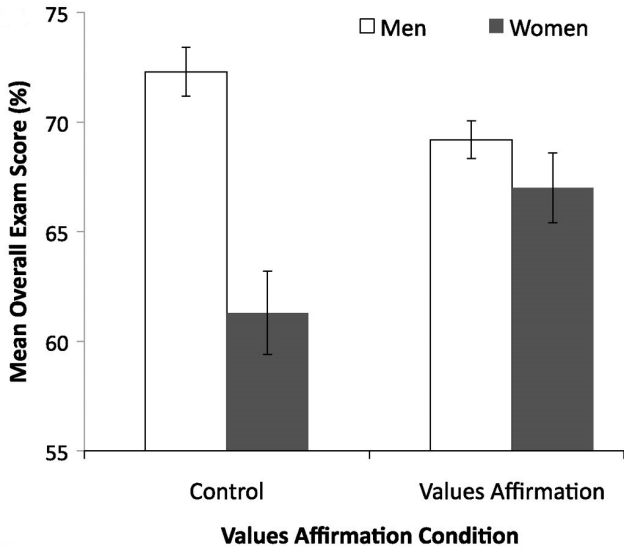
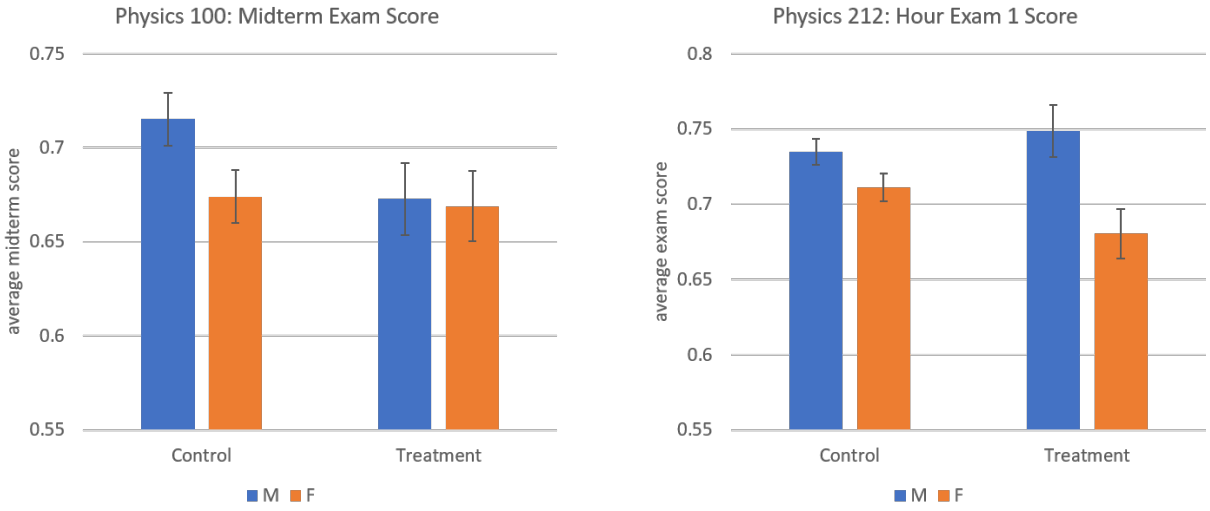


Figure 11.1: Adjusted exam scores for students in the Miyake et al study, by gender and treatment condition [192]. Note that the vertical axis begins at 55%.

the difference between Treatment and Control women. We did not do adjustments in our analysis. Our raw exam scores for students in each group is shown in Figure 11.2. We saw virtually no difference in women's scores in Physics 100, regardless of treatment, and lower performance by men in the treatment group. These changes slightly reduced the gender gap by about 4%, but did not elevate women's scores and the change was largely due to decline in men's scores with the treatment. In Physics 212, women who received the treatment did significantly worse and men with the treatment did slightly better, which increased the gap to about 7% for the treatment condition compared to 2% in control.



(a) (b)

Figure 11.2: Average exam scores for men and women in treatment (values affirmation) and control conditions for (a) Physics 100 and (b) Physics 212. Note that the vertical axis begins at 0.55.

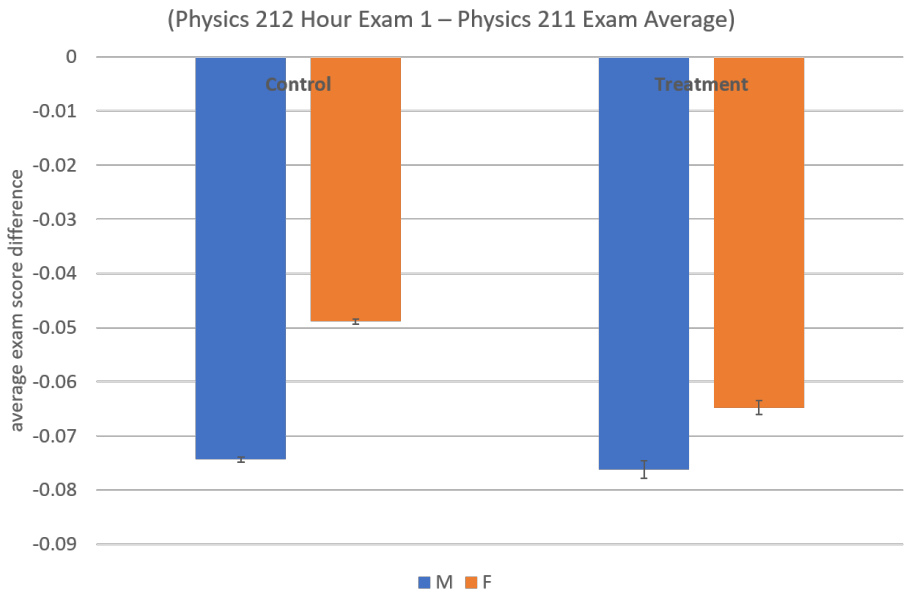


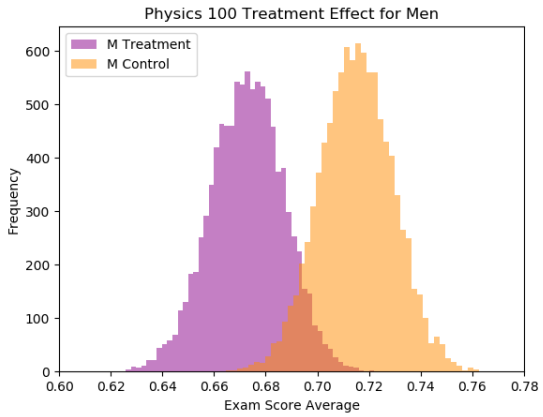
Figure 11.3: Average scores on 212 exam with 211 exam scores subtracted out for each group.

For students in Physics 212, we could adjust for students' prior exam grades in Physics 211, the preceding course. By plotting students' Physics 212 exam score differences from their 211 exams, we have a stronger result which controls for differences among student groups; men performed about equally on their exams, and women still performed worse in the treatment condition. This comparison does better minimize the gap by equalizing the differences in men, but increases the difference between women in the treatment and control groups, opposite of the way the intervention is intended to work.

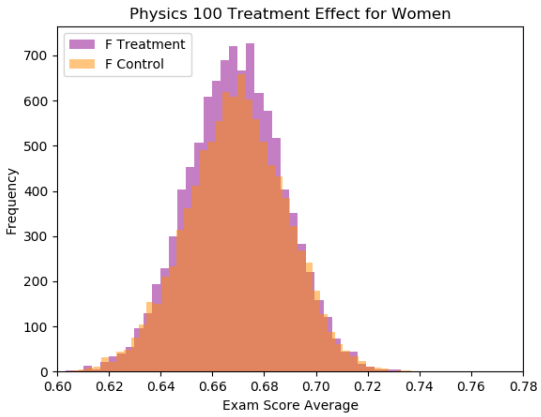
Bootstrap Analysis

Students' scores were bootstrapped by randomly selecting students from within each condition (gender x treatment) to simulate 10,000 trials with our data. Histograms of the distributions of bootstrapped average exam scores is shown in Figure 11.4. A quick look at these plots shows us that the only group that saw improved exam performance due to the treatment was men in Physics 212. Women in Physics 100 saw effectively no change from the affirmation exercise, while men in Physics 100 and women in Physics 212 saw a negative effect from the treatment.

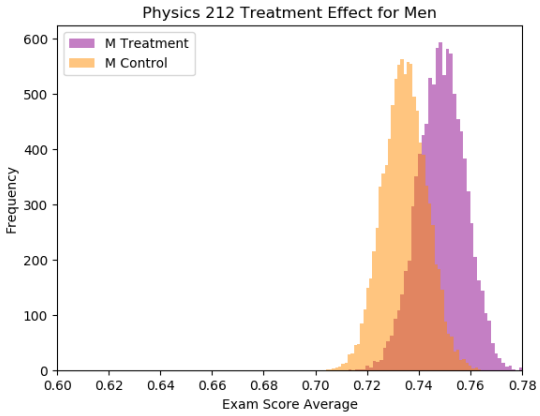
Each of these differences have their own distributions, which are shown in Figure 11.5. According to theory, the affirmation activity should help women more than men, so a plot showing the effects for women, subtracting the effects for men is also included in Figure 11.6. We primarily considered the effects within gender, rather than the comparison of men to women, but this plot shows contradictory effects in the two courses. Within each gender group, a calculation of the mean average gains for students with the treatment over students with the control condition are shown in Figure 11.7 with 95% confidence intervals from the bootstrap as error. For completeness and comparison, this plot also includes the adjusted scores reported by Colorado in its two implementations on exams and FMCE gains, with their standard error bars transformed into confidence intervals [192, 256].



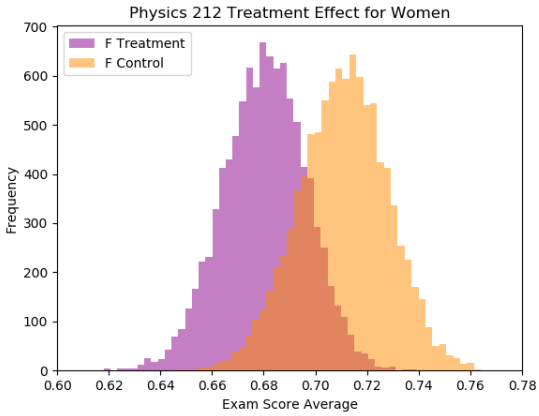
(a) Men in Physics 100



(b) Women in Physics 100

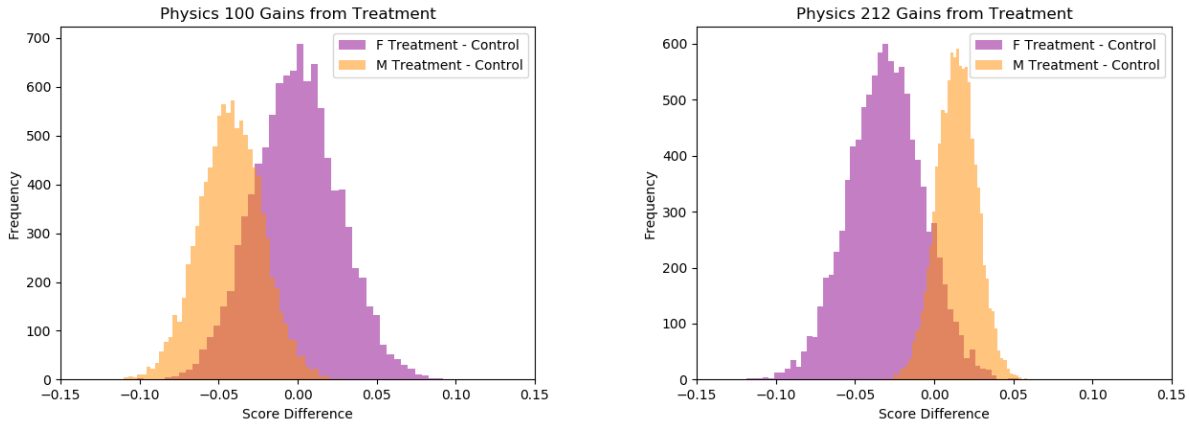


(c) Men in Physics 212



(d) Women in Physics 212

Figure 11.4: Histogram of bootstrapped exam average ranges for men and women in treatment (values affirmation) and control conditions in Physics 100 and 212. In all of these plots, the control condition is light orange and the treatment condition is magenta; dark orange shows their overlap.



(a) Physics 100

(b) Physics 212

Figure 11.5: Histogram of bootstrapped exam differences for men and women in (a) Physics 100 and (b) Physics 212. These gains were calculated by subtracting the average exam scores for those in the control condition from exam scores for those in the treatment condition, meaning positive gains imply higher scores from the affirmation exercise while negative gains show lower scores in those who completed the affirmation.

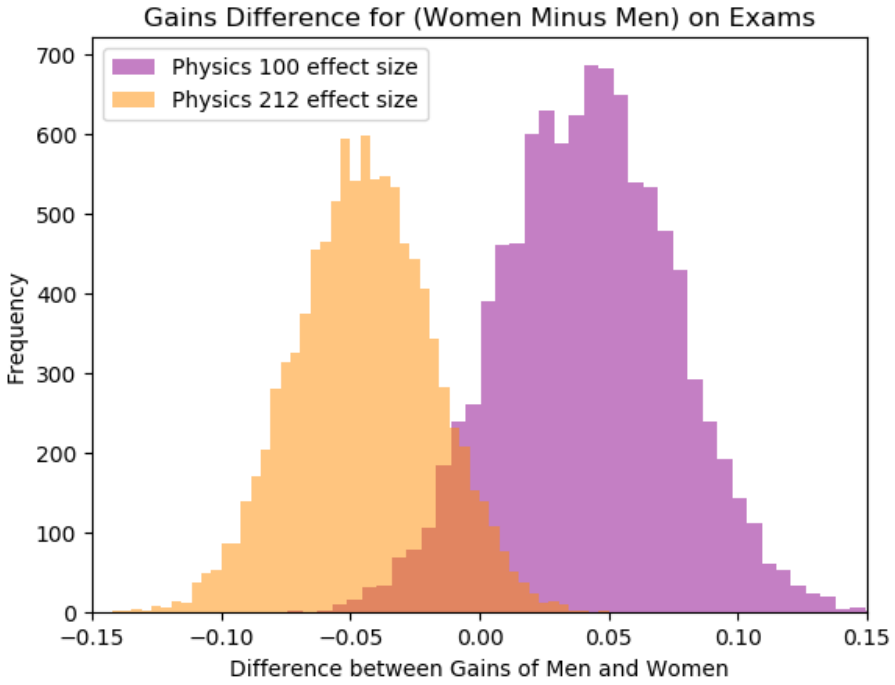


Figure 11.6: Histogram of boostrapped exam gains for women over men due to the affirmation activity. These values are the subtraction of gains by men from the gains by women, effectively the series in each plot of Figure 11.5 subtracted from each other.

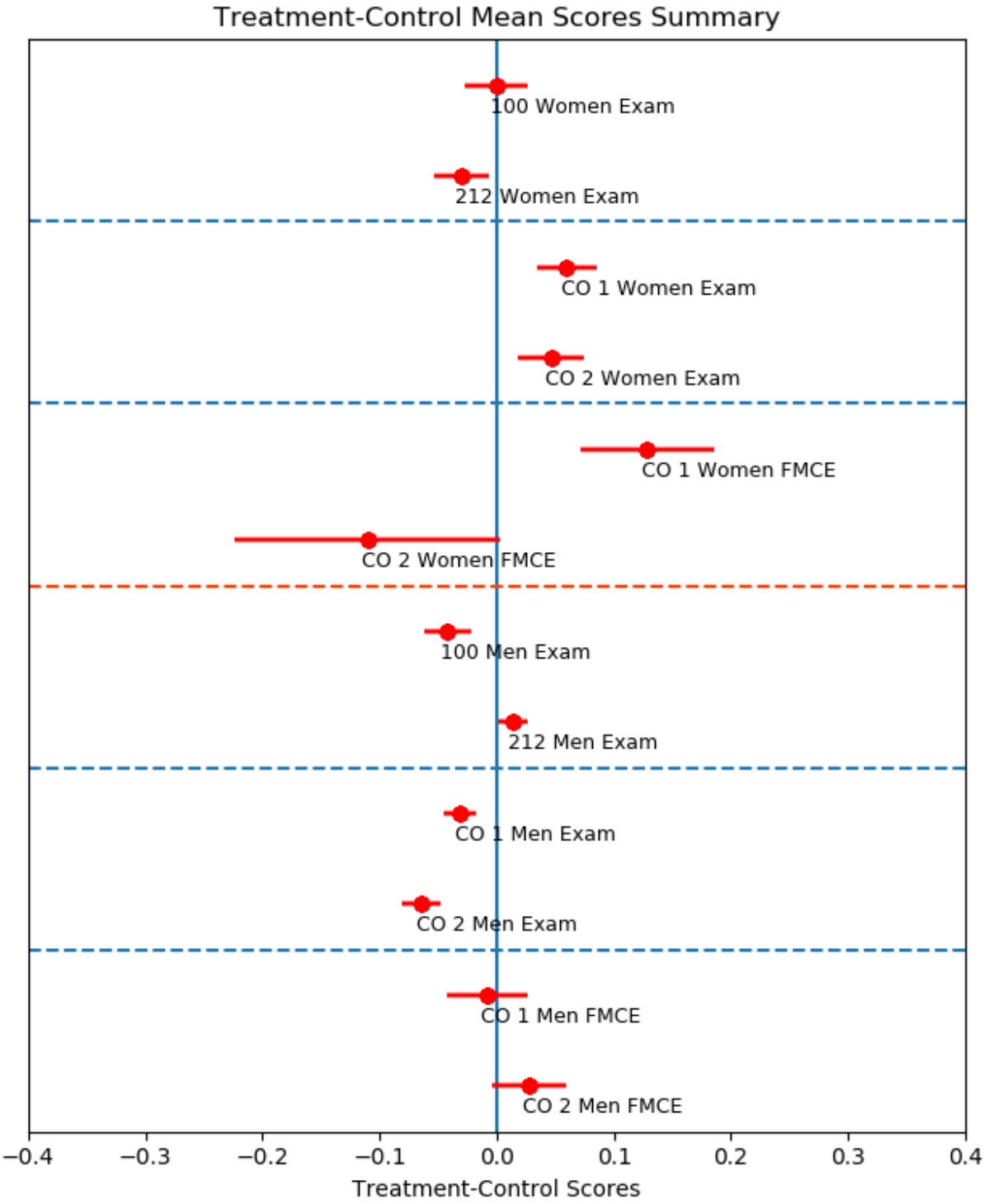


Figure 11.7: Summary of effects within each gender and treatment. Points indicate mean value of Treatment - Control values and error bars are standard error. CO 1 refers to the first Colorado experiment [192] and CO 2 refers to the second implementation by Colorado [256]. The values for Colorado are adjusted scores, and their error bars are their reported standard errors.

11.3.2 Final Grades

The researchers at Colorado also investigated the effect of the values affirmation exercises on students' final grades and saw an increase in the number of As, Bs, and reduction in number of Cs for women in the treatment group compared to women in the control group [192]. For reference, their plot is shown in Figure 11.8. The same analysis on our population showed the opposite effect; both courses saw a larger percentage of women in the control condition earning As than in the treatment, though these were the only differences in distribution that neared significance. These distributions are shown in Figure 11.9; the values were not boostrapped as the Colorado paper did not include error and could not be compared.

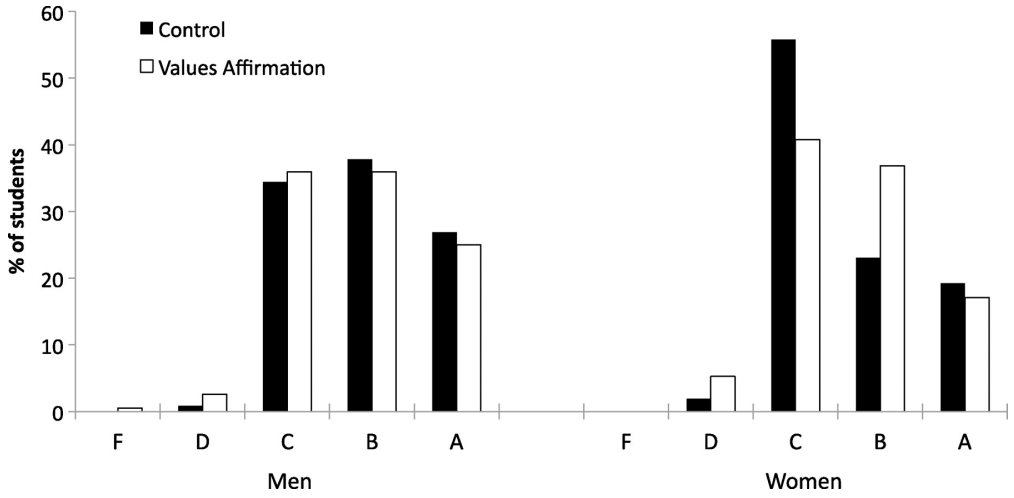
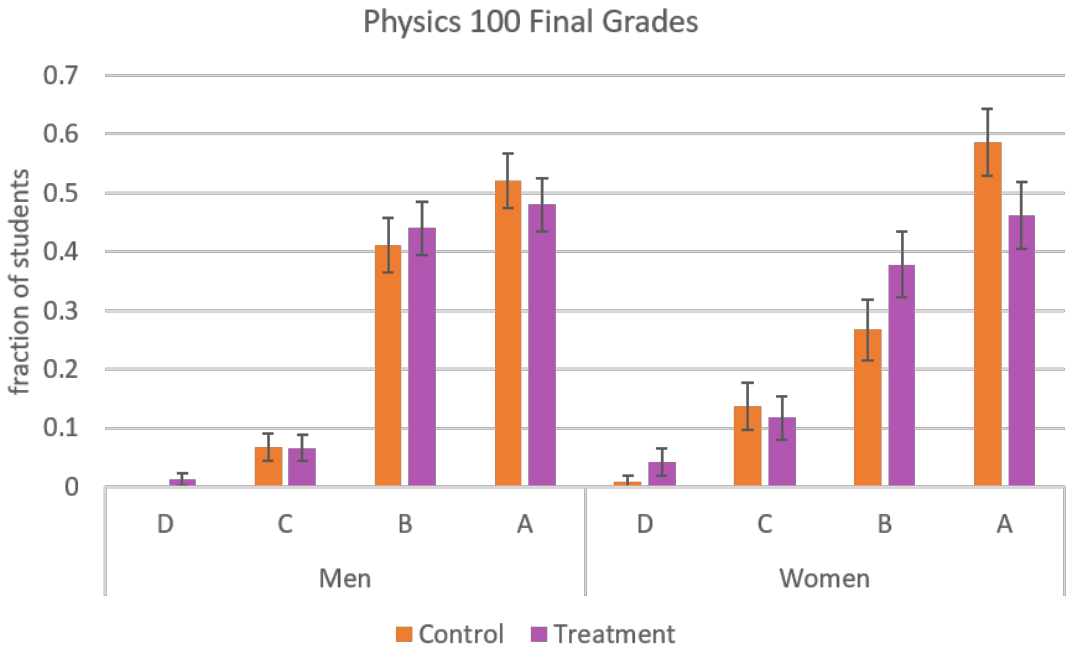
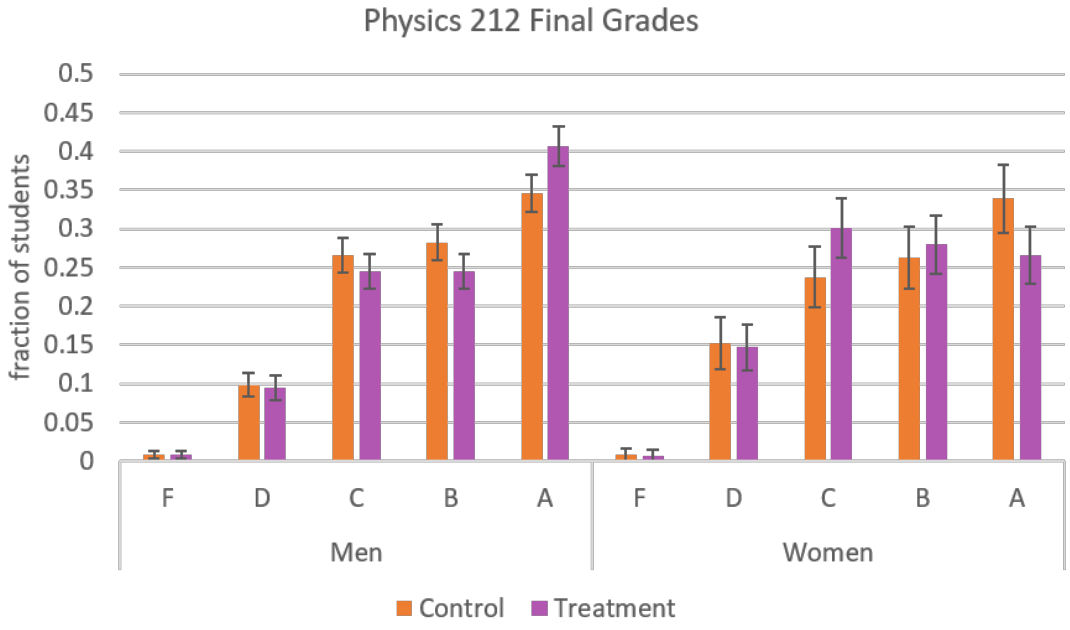


Figure 11.8: Final grades of students by gender and treatment or control condition in the original Colorado implementation [192].



(a) Physics 100



(b) Physics 212

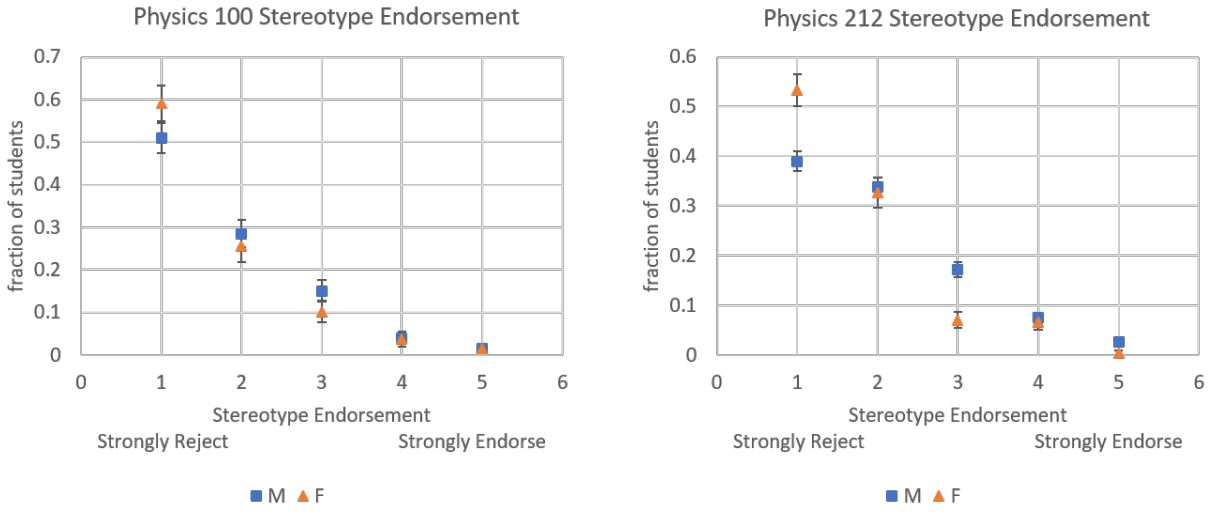
Figure 11.9: Distributions of students' final grades within each gender and treatment condition for (a) Physics 100 and (b) Physics 212. Each of these fractions is calculated out of the total number of students in each category.

11.3.3 Stereotype Endorsement Effects

Raw Data

A major result from the Colorado paper we attempted to replicate was the increased gains for women who endorse gender stereotypes in physics compared to those with a low endorsement [192]. For our analysis, we used the same Likert scale agreement survey statement: "According to my own personal beliefs, I expect men to generally do better in physics than women." In the following plots, students answering 1 (one) corresponds to "Strongly Disagree," a strong rejection of the stereotype, through 5 (five) corresponding to "Strongly Agree," a strong endorsement of the stereotype. Students also had the option to choose "No Response," which was selected by about 1% of students in each course, and those students are not included in the following analysis. To get a snapshot of what students' level of endorsement is, a scatter plot of the fraction of students answering each choice from 1 (strong rejection) through 5 (strong endorsement) is given in Figure 11.10.

To classify our students as high and low endorsers of the stereotype, we used the same definitions as Miyake et al [192]: low endorsers were students who answered below 0.75 times the standard deviation below the mean, and high endorsers were students who answered 0.75 times the standard deviations above the mean, treating students' responses as a continuous variable. In Physics 100, the mean value of students' responses was 1.71, with a standard deviation of 0.94. The mean value for Physics 212 responders was 1.92 with a standard deviation of 1.0. In both courses, the cutoff for "high" endorsers was any response of 3 (neutral) or higher, and only responses of 1 corresponded to "low" endorsers. Table 11.3 gives the number of students in each condition using this categorization; these students are a subset of the original population as students who responded "2: Disagree" and "No Response" are not included in either group, nor are the students who



(a) Physics 100

(b) Physics 212

Figure 11.10: Fraction of men and women responding for each level of endorsement to the statement, "According to my own personal beliefs, I expect men to generally do better in physics than women." Students selecting 1 "strongly disagree" with the statement (a rejection of the stereotype), while students selecting 5 "strongly agree" with the statement (an endorsement of the stereotype).

Table 11.3: Number of participants in each condition, by course and stereotype endorsement.

		Physics 100		Physics 212	
		Treatment	**Control**	**Treatment**	**Control**
Low Endorsing	**M**	43	59	127	119
	F	40	41	73	56
High Endorsing	**M**	22	19	83	88
	F	9	12	18	16

did not complete the survey.

From their classification, Miyake et al saw dramatic results, shown in Figure 11.11. Our results, presented in the same way, are given in Figure 11.12. In Physics 100, we see again that changes in the gap are predominantly due to men in the treatment group underperforming men in the control group. We do see higher scores for high endorsing women with the affirmation treatment in Physics 212 accompanied by lower scores for low endorsing women with treatment. However, these differences are within error bars and the representation of these groups as progressions is misleading; each point is a separate population of students and the format was copied for completeness in gauging replication. A better comparison lies in looking at differences within like groups.

The same results are plotted as a bar chart in Figure 11.13 which allows the reader to compare students with high or low endorsement within each group. The only statistically significant difference between high and low endorsing students is seen in women in the control group for Physics 212; women in the control group who endorse the stereotype did 10% worse on average on their exam than women who reject the stereotype. This difference was not present in women who participated in the treatment, which may suggest that the treatment protected high endorsing women from the effects of the stereotype. One can also measure the effects of the affirmation within low and high endorsing men and women, which is shown in Figure 11.14. Again, we see some benefit for high endorsing women in Physics 212, but the low endorsing students did worse with the affirmation which amplifies the difference seen in Figure 11.13.

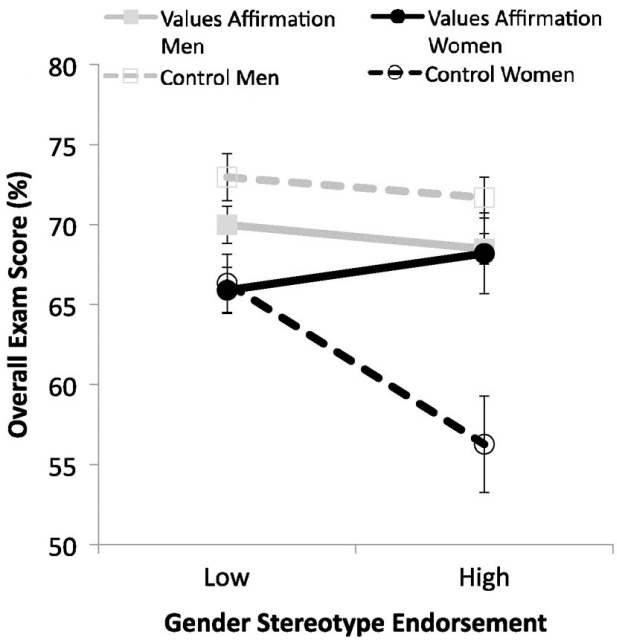
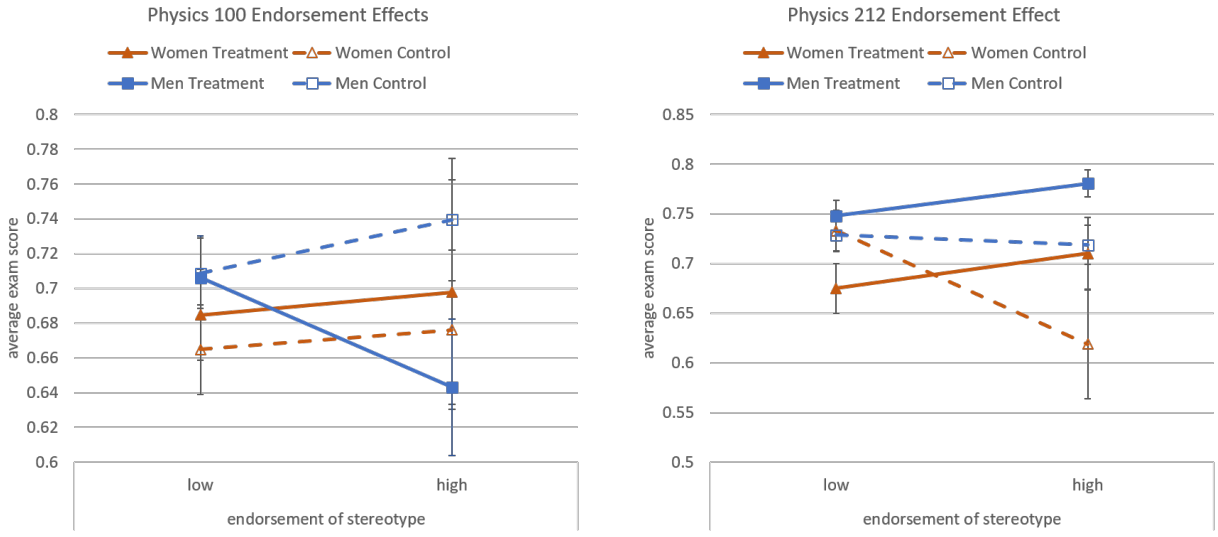


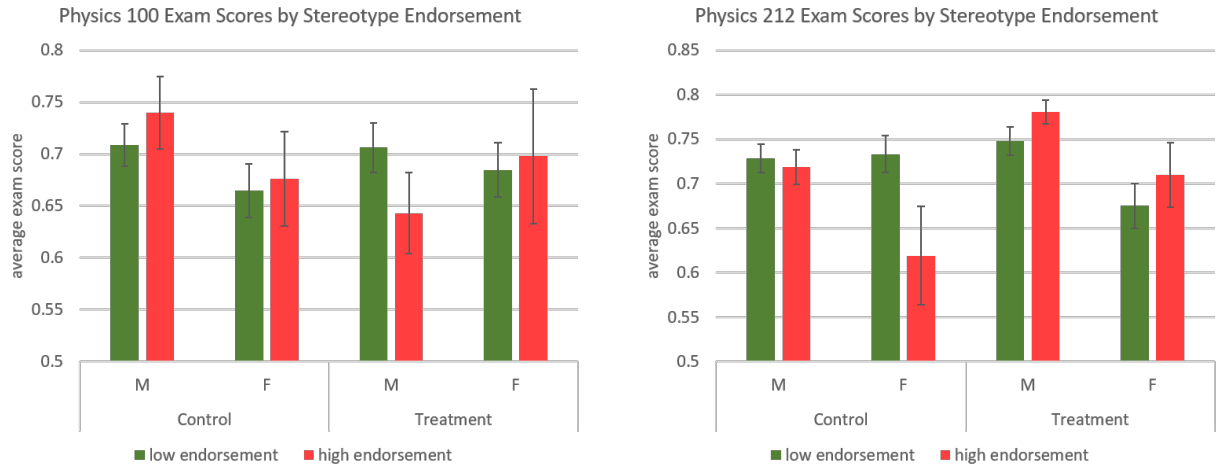
Figure 11.11: Students' exam scores by treatment and gender group, also partitioned by high and low stereotype endorsement, for the first Colorado implementation [192].



(a) Physics 100

(b) Physics 212

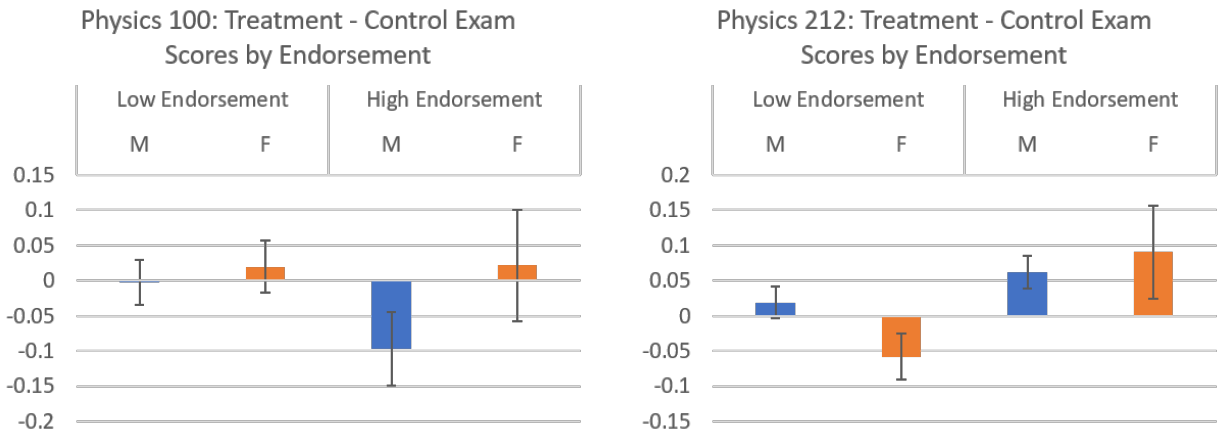
Figure 11.12: Average exam grades for men and women in each treatment group, based on their classification of being a high or low endorser of gender stereotype in physics.



(a) Physics 100

(b) Physics 212

Figure 11.13: Average exam grades for men and women in each treatment group, based on their classification of being a high or low endorser of gender stereotype in physics.



(a) Physics 100

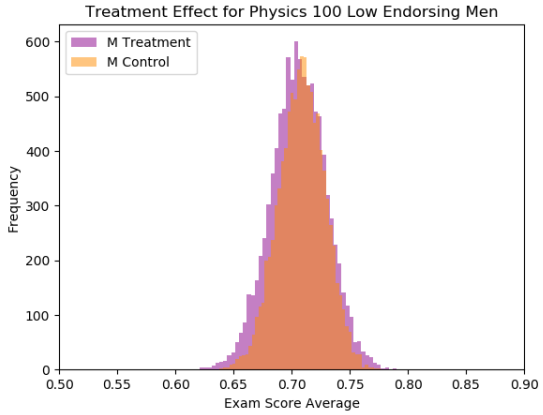
(b) Physics 212

Figure 11.14: Difference in average exam grade for students with the treatment condition compared to students with the control condition, based on their level of gender stereotype endorsement.

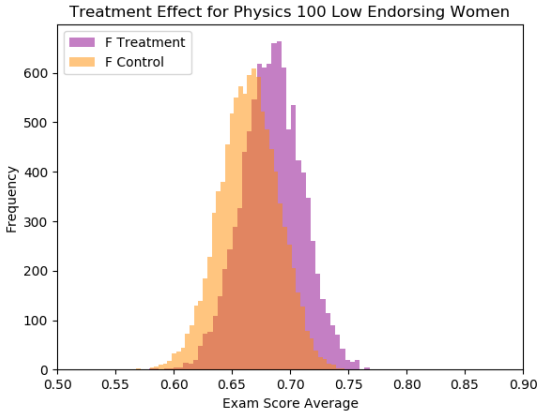
Boostrap Analysis

Students' scores were again bootstrapped by randomly selecting students from within each condition (gender x treatment x endorsement) to simulate 10,000 trials. The distributions of average exam scores for each subgroup, comparing average exam scores of students with treatment and control condition and separated by gender stereotype endorsement are shown in Figure 11.15 (Physics 100) and Figure 11.16 (Physics 212).

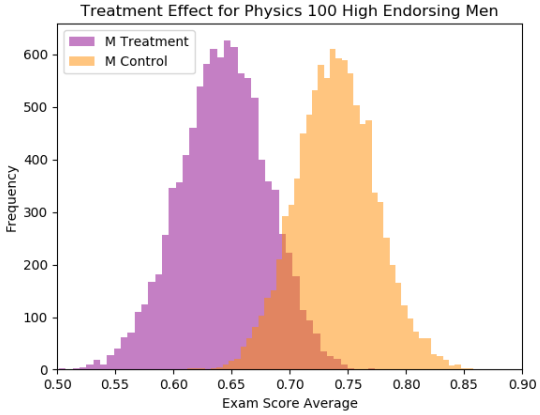
Like in the previous bootstrap analysis, these groups can be subtracted from each other to see the effects of the affirmation treatment. Each of the plots in Figure 11.17 show the control students' average scores subtracted from the treatment students' scores, or the gains from students who participated in the affirmation activity. A summary of these difference distributions is shown in Figure 11.18, with bootstrapped 68% confidence intervals given as error bars (a standard error). As with the last summary, values from the first Colorado implementation [192] are included on the plot for comparison; the second Colorado implementation did not include information about stereotype endorsement.



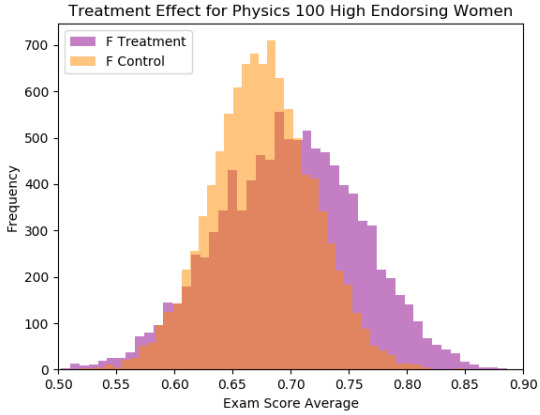
(a) Men with low gender stereotype endorsement



(b) Women with low gender stereotype endorsement

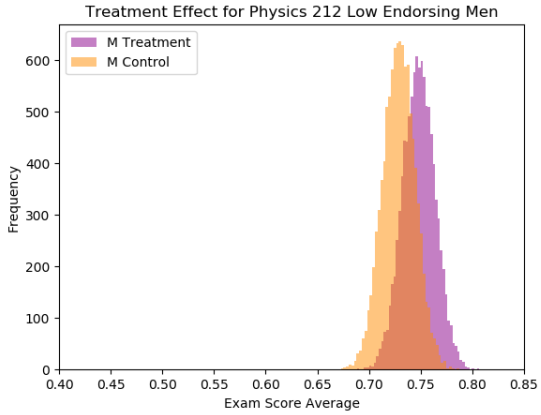


(c) Men with high gender stereotype endorsement

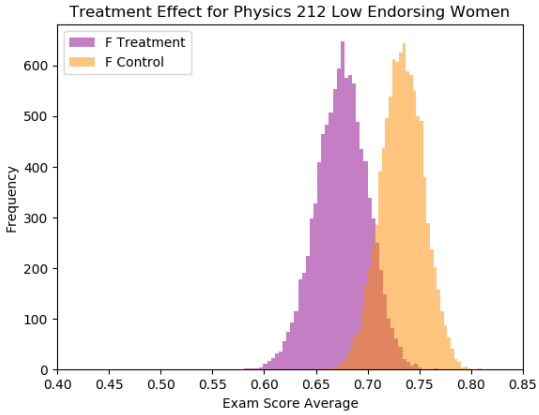


(d) Women with high gender stereotype endorsement

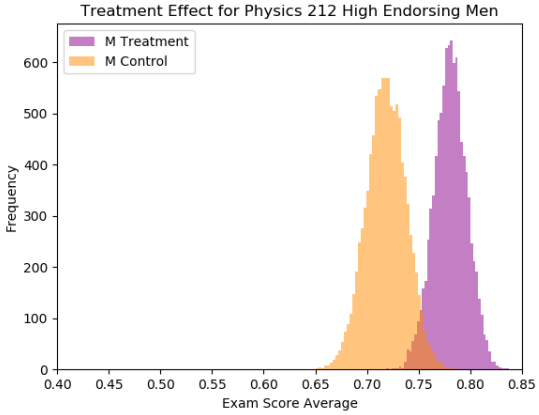
Figure 11.15: Histogram of bootstrapped exam average ranges in Physics 100 for men and women in treatment (values affirmation) and control conditions with low and high endorsement of gender stereotype in physics. In all of these plots, the control condition is light orange and the treatment condition is magenta; dark orange shows their overlap.



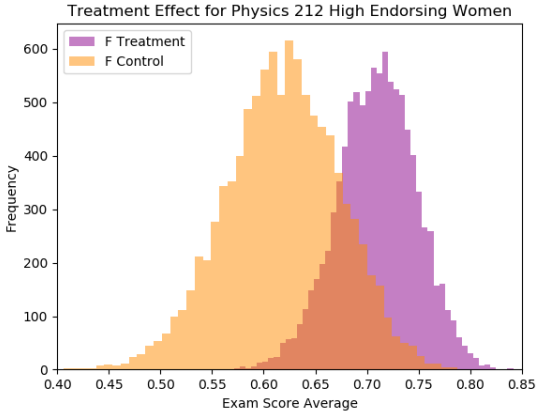
(a) Men with low gender stereotype endorsement



(b) Women with low gender stereotype endorsement

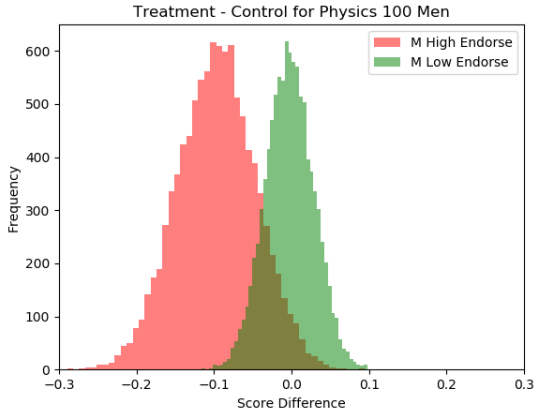


(c) Men with high gender stereotype endorsement

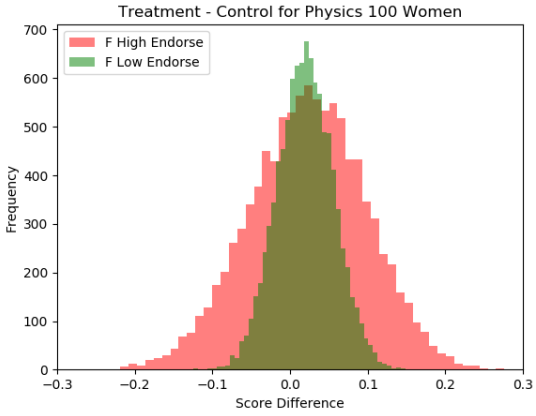


(d) Women with high gender stereotype endorsement

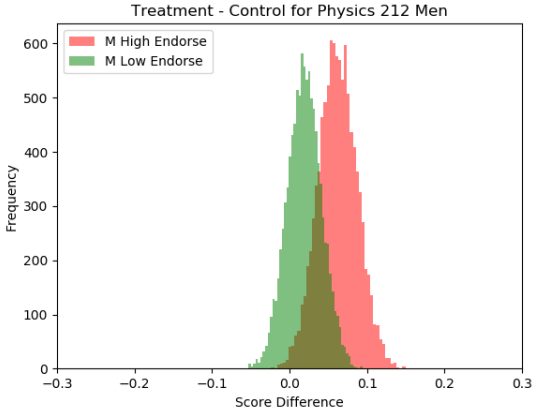
Figure 11.16: Histogram of bootstrapped exam average ranges in Physics 212 for men and women in treatment (values affirmation) and control conditions with low and high endorsement of gender stereotype in physics. In all of these plots, the control condition is light orange and the treatment condition is magenta; dark orange shows their overlap.



(a) Physics 100 Men



(b) Physics 100 Women



(c) Physics 212 Men

(d) Physics 212 Women

Figure 11.17: Histogram of bootstrapped exam gains for treatment participants over control participants, comparing low and high stereotype endorsement. These gains were calculated by subtracting the average exam scores for those in the control condition from exam scores for those in the treatment condition, meaning positive gains imply higher scores from the affirmation exercise while negative gains show lower scores in those who completed the affirmation. In these plots, high endorsers are plotted in red and low endorsers are plotted in light green; dark green shows their overlap.

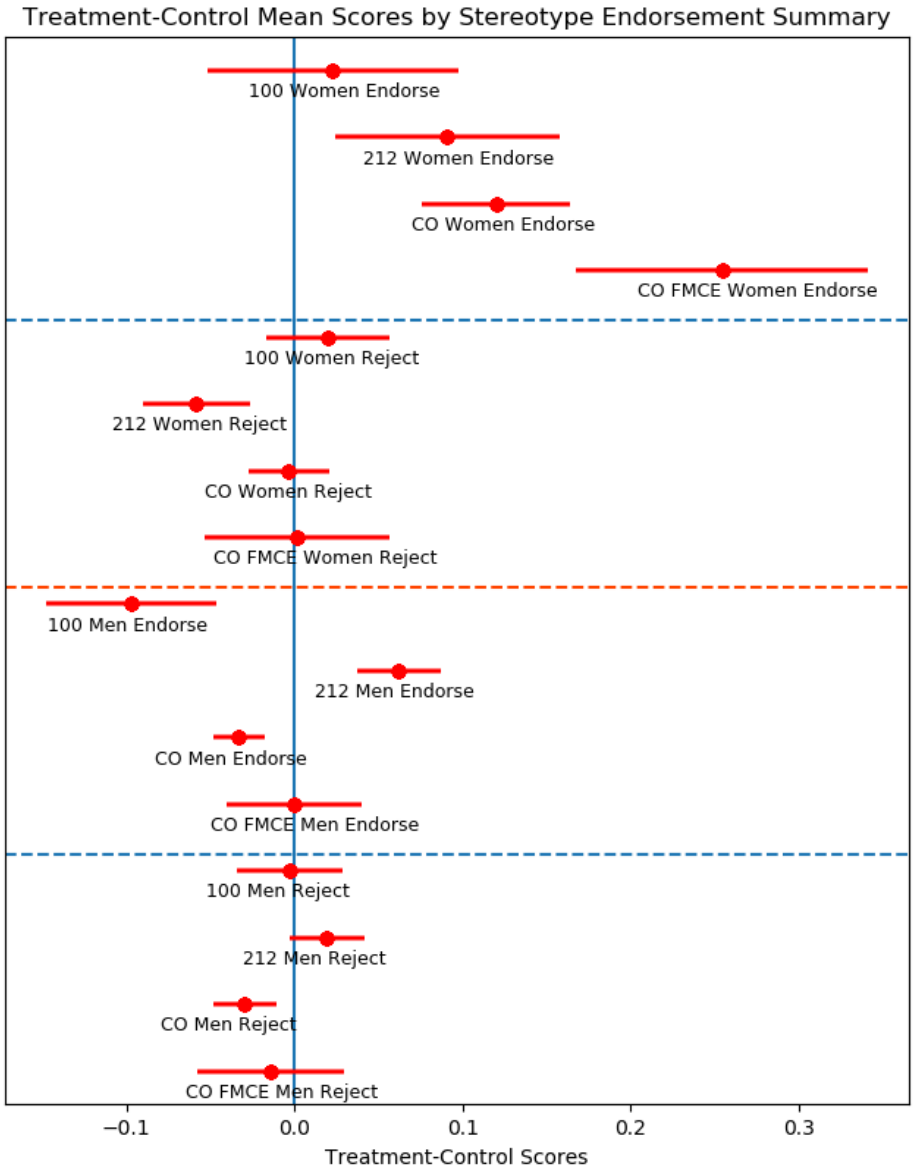


Figure 11.18: Summary of effects within each gender and treatment, partitioned by stereotype endorsement. "Endorse" values refer to student averages of those who had high endorsement of the stereotype, while "Reject" values refer to low endorsement. Points indicate mean value of Treatment - Control average values and error bars are standard error. CO refers to the first Colorado experiment [192]. The values for Colorado are adjusted scores, and their error bars are their reported standard errors.

11.4 Discussion

In consolidating all our data and the data from Colorado, we primarily consider the effects of the affirmation on women; we argue that it is more important to lift women's scores which may be affected by stereotype threat than to suppress men's scores to minimize the gap. In the original Colorado studies [192, 256], some of the reduction in their gender gaps were due to lower men's scores in the treatment group, which we also saw in our results for Physics 100. The gap "decreased" due to lower men's performance in the treatment group although women's performance did not change from the treatment. Woolf et al saw a similar "decrease" in gaps via White students performing worse and those from ethnic minority groups being unaffected [239]. These results exemplify some of the dangers of measuring gaps to identify success; setting the majority group as the "ideal" is not productive if the intervention is lowering their performance, and pitting the achievements of majority and minority groups against each other does not necessarily elevate minoritized students' abilities.

When considering gains by women at the University of Illinois, our two results combine to give us a most likely value of the effect of the treatment for women to be -1.8 $\pm$ 3.5% (where error is a 95% confidence interval). In comparison, the Colorado papers saw an average effect of 5.8 $\pm$ 2%. If we combine all our data, the value which best represents the effect is 2 $\pm$ 2%, but one can see from the summary in Figure 11.7 that the spread of the data is larger than one would expect, so there are likely other factors influencing the results. Our analysis and confidence intervals at least set an upper boundary on what one might expect in the context of a predominantly White Research 1 University.

There are many effects which contribute to differences in performance between men and women in physics classrooms, but if we isolate the effects of stereotype threat, the intervention should be most useful for women who endorse gender stereotypes in STEM. The threshold which we used to determine "high" and "low" endorsers is somewhat arbitrary but was borrowed from Colorado's implementation. Initially, including "Neutral" responses (response 3) to the statement than men perform better than women as "high" endorsing felt strange, but further reflection has made us more comfortable with this categorization; if a person does not outright reject the notion that men perform better than women, this is a form of endorsement. The survey item is not suggesting that women must perform better than men for students to disagree with the statement, just to assert whether they believe women inherently perform worse.

When we restricted our analysis to only students who marked their survey responses with 4 or 5, corresponding to "agreeing" and "strongly agreeing" that men perform better than women, our sample sizes were very small and showed no statistically significant improvements. In particular, women with the treatment who endorsed by "agreeing" (a response of 4) saw a decline in performance and both high endorsing men and women in Physics 212 saw the same improvement, which should not be true via the affirmation theory. Both courses did not have women "strongly agreeing" (a response of 5) in both treatment and control groups to even calculate the difference between the conditions. Although not elaborated on in this chapter, students' responses to societal expectations showed much more variation, and believing that society expects you to fail whether you personally subscribe to the stereotype may also affect students' susceptibility to stereotype threat.

Using the more generous definitions of low and high endorsers for our analaysis, we saw that the effects for high endorsing women averaged around 6 $\pm$ 3% at the University of Illinois, where again error cited is

a 95% confidence interval. At Colorado, their two reported effects can be combined to estimate an 15 $\pm$ 8% improvement, though they did not report data about stereotype endorsement in their replication. Their largest effect in the original study was seen in high endorsing women on the FMCE, and their replication saw that generally women with the treatment performed much worse on the FMCE than those in control, so it is possible that a null or negative result regarding stereotype endorsement was not included in their write up of the replication. When we combine all results from the Universities of Illinois and Colorado, the most statistically likely result is an 11 $\pm$ 7% improvement for high endorsing women. However, the value is driven up by Colorado's large result of 25% improvement for their endorsing women on the FMCE, which is statistically inconsistent with all other results, including their own, and was not reported at all in their followup study.

When results are presented from individual studies, it is more likely that a statistically significant fluctuation may be interpreted as positive results, but combining multiple studies together diminishes the strength of those results. A takeaway from our disparate results is that the variables affecting student performance are not as simple as doing the activity versus not doing the activity. It is likely that additional factors contribute to the effect which are not explicitly defined or described; following the protocol outlined by the Colorado paper was not sufficient to replicate success, but further studies may better illuminate those hidden variables.

Chapter 12

Conclusions: Implementation and Affect Matter

In the work described, we created tools in an attempt ease students' transition from disparate high school physics backgrounds into our physics engineering sequence. The largest takeaway from the experience has been our results' emphasis on the entanglement of student affect and the effectiveness of the tools. In many instances when theory would predict large experimental results, student frustration overshadowed potential benefits; the implementation of these tools, however well-intended, is sensitive and can be dominated by student frustration, so they must be created and used with attention to students' emotional response.

After promising results in our mastery clinical trials, which had volunteers use mastery exercises for a single afternoon, the same results did not translate to the semester-long implementation of mastery homework. Groups of students behaved very differently in the freshman preparatory course than in a single implementation for a more advanced course. Although during the clinical trials, some students showed outward signs of frustration, this was amplified and prolonged for students who used the mastery homework for the entire semester. This frustration was documented via online surveys and can explain many non-productive behaviors, such as skipping through intermediate versions and low engagement with solution videos. From the open-ended responses to the survey, students revealed a lacking sense of agency while using the system. Many believed the homework to be unfair and punishing, an obstacle to be overcome rather than a tool which was intended to help them.

After attending to these frustrations via a combination of content restructuring, a change in delivery method, and a homework priming exercise, students' engagement with the mastery system improved considerably Self-reported levels of frustration were cut in half and students spent more time with solutions, spoke of the system more positively, and performed better on the exercises. By allowing students to retain points even without mastering, placing the decision to keep working in their control, more students successfully mastered the same levels.

Even with these changes, we discovered that many levels were too difficult for students to master and we tried different approaches to improve student success. By including more scaffolding to a notoriously challenging level, students were able to perform better on the mastery levels themselves and on followup activities with

the same amount of time on task. Similarly, by breaking levels into smaller pieces with more specific content and competencies, students were able to spend more of their time on the competencies they had not yet mastered. Repetition of problems they understood was a frequently cited source of frustration, and ensuring that the levels were not too broad in scope could alleviate some of that frustration. One concern in making levels more specific or providing extra scaffolding is that treating content as too disparate or providing too much guidance may hurt students' performance when faced with a whole problem without help. Our these studies saw that the treatments did not hurt overall performance or the ability to synthesize when the help was removed.

Similarly, our implementation of bi-weekly quizzes and retakes for Physics 100 students showed the importance of careful implementation. In our first year of bi-weekly quizzes, no penalty for doing worse on the retakes may have limited their potential. We saw improvement in that students' retake scores tended to regress towards the mean and we saw fewer students failing the exam, but we also saw students using retakes even when they had high scores on the original quiz and students leaving retake exams partway through, as there was no penalty for a low retake score. Much of the theory behind frequent testing relies on students' incentive to study between quizzes, and having no stakes associated with the retake may have suppressed potential gains by allowing students to use retakes even if they had no opportunity for further study between the quiz and retake.

When exam retakes were implemented in Physics 211, the following course, an equal weight of the initial and retake exam successfully refined the population using retakes to be mostly low-scoring students as intended. Their retake exam scores improved but much can be explained by a regression towards the mean, rather than actual learning gains, as low-scoring students in a year without retakes saw similar gains from one exam to the next. Even the effects of practice exams were sensitive to their delivery. when the practice exams were given as unlimited attempt exercises alongside solution videos, student engagement with the practice exams had no correlation to their actual exam scores and often inflated students' perceived competence by allowing multiple attempts and providing help which would not be on the exam. When the practice exams were changed to allow students only a single submission and forced them to make that submission before accessing the help video, students saw improvement within the practice exams themselves and a better correlation to exam scores.

Our attempted replication of a values affirmation intervention was slightly disconnected from our other efforts, but the emphasis on student affect learned from our first experiments made the exercise appealing as a way to focus on the social aspects underlying students' emotions directly. In Physics 100 specifically, we saw that the treatment hurt women at large, though not statistically, and our results again caution against assuming theory will necessarily translate to results will not always be a correct assumption. Especially when dealing with something as complex as students' identities and that interplay with their behavior and performance, there is not a quick and generalizable solution. What works in one context may completely fail in another.

We have seen that the best-motivated interventions may fail due to sensitivity of their implementation and student affect and that iterative improvements with attention to these details are key to success. Mastery-style exercises showed learning gains when some of the obstacles to student engagement were addressed, and frequent testing improved student grades, just most effectively when given in a practice exam which mimicked an actual exam. By being responsive to our results and adjusting implementation iteratively, we

were able to amplify their effects and refine our understanding of the contributing mechanisms.

Moving forward with this work, there are still areas to improve and explore. Within our mastery exercises in Physics 100, many levels persist which have very low mastery rates, and addressing these difficult levels provides more opportunity to learn the limits of mastery-style exercises or continue to develop ways to make them work for difficult or computation-heavy material. The experiment documented in Chapter 8 which tested the optimal grain-size of content in levels will be repeated with more homogenous content between "split" and "whole" levels, which may provide a stronger result. Further, looking at students' progression between tries can give us information about how students work through different versions and how well answering a question correctly on one version can translate to another; the idea that correctly answering questions translates to mastery of the material is a core assumption of our homework.

Beyond the direct scope of the work in this thesis, I am particularly interested in interventions and factors which may foster students' sense of ownership and self-efficacy, a major theme which emerged from our course-wide implementation of mastery homework. Particularly for students who are at higher-risk for feeling a lack of belonging in STEM, I believe that systems which provide positive feedback frequently, are adaptive to students' needs, and draw on students' owned experiences are ideal for engaging students, as we have tried to do. However, I would also like to work to more formally understand, document, and explore engagement mechanisms from a more holistic framework and existing cognitive literature to learn how to better achieve these goals. In instances where well-intentioned learning design does not resonate with students' incentives, more attention to students' experiences are necessary, and focus on methods to nurture student ownership of learning may help to deliberately design to align activities with students' ethos. The perceived tension for instructors between pushing students to succeed and meeting them at their current state does not need to be oppositional; with targeted strategies, both goals can be achieved.

References

[1] Homeyra R. Sadaghiani. "Using multimedia learning modules in a hybrid-online course in electricity and magnetism". In: *Physical Review Special Topics-Physics Education Research* 7.1 (2011), p. 010102.

[2] Homeyra R. Sadaghiani. "Online prelectures: An alternative to textbook reading assignments". In: *The Physics Teacher* 50.5 (2012), pp. 301–303.

[3] Homeyra R. Sadaghiani. "Controlled study on the effectiveness of multimedia learning modules for teaching mechanics". In: *Physical Review Special Topics-Physics Education Research* 8.1 (2012), p. 010103.

[4] Zhongzhou Chen, Timothy Stelzer, and Gary Gladding. "Using multimedia modules to better prepare students for introductory physics lecture". In: *Physical Review Special Topics-Physics Education Research* 6.1 (2010), p. 010108.

[5] Timothy Stelzer, Gary Gladding, José Mestre, and David T. Brookes. "Comparing the efficacy of multimedia modules with traditonal textbooks for learning introductory physics content". In: *American Journal of Physics* 77.2 (2009), pp. 184–190.

[6] Timothy Stelzer, David T. Brookes, Gary Gladding, and José P. Mestre. "Impact of multimedia learning modules on an introductory course on electricity and magnetism". In: *American Journal of Physics* 78.7 (2010), pp. 755–759.

[7] Benjamin S. Bloom. "Learning for Mastery." In: *Evaluation Comment* (1968).

[8] Thomas Guskey. "Closing Achievement Gaps: Revisiting Benjamin S. Bloom's "Learning for Mastery"". In: *Journal of Advanced Academics* 19 (2007), pp. 8–31. DOI: 10.4219/jaa-2007-704.

[9] John B. Carroll. "A Model of School Learning". In: *Teachers College record* 64 (1963), pp. 723–733.

[10] James H. Block and Robert B. Burns. "1: Mastery Learning". In: *Review of Research in Education - REV RES EDUC* 4 (1976), pp. 3–49. DOI: 10.3102/0091732X004001003.

[11] Brianne Gutmann, Gary Gladding, Morten Lundsgaard, and Timothy Stelzer. "Mastery-style homework exercises in introductory physics courses: Implementation matters". In: *Phys. Rev. Phys. Educ. Res.* 14.1 (2018), p. 010128. DOI: 10.1103/PhysRevPhysEducRes.14.010128.

[12] Carleton W. Washburne. "Educational Measurement as a Key to Individual Instruction and Promotions". In: *The Journal of Educational Research* 5.3 (1922), pp. 195–206.

[13] Henry Clinton Morrison. *The practice of teaching in the secondary school*. Chicago, Ill. : The University of Chicago press, 1926.

[14] K. Spencer. *Media and Technology in Education: Raising Academic Standards*. Manutius Press, 1996.

[15] B. F. Skinner. *The Science of Learning and the Art of Teaching*. Harvard Educational Review, 1954.

[16] B. F. Skinner. "Why We Need Teaching Machines". In: 31 (1961), pp. 377–398.

[17] F. S Keller. "Good-bye, teacher". In: *Journal of applied behavior analysis* 1 (1968), pp. 79–89.

[18] Paul Black and Dylan Wiliam. "Assessment and Classroom Learning". In: *Assessment in Education* 5 (1998), pp. 7–74. DOI: 10.1080/0969595980050102.

[19] Paul J. Black. "Formative and Summative Assessment by Teachers". In: *Studies in Science Education* 21.1 (1993), pp. 49–97. DOI: 10.1080/03057269308560014. eprint: https://doi.org/10.1080/03057269308560014.

[20] D. Royce Sadler. "Formative assessment and the design of instructional systems". In: *Instructional Science* 18.2 (1989), pp. 119–144.

[21] Philippe Perrenoud. "From formative evaluation to a controlled regulation of learning processes. Towards a wider conceptual field". In: *Assessment in Education: Principles, Policy & Practice* 5.1 (1998), pp. 85–102.

[22] Beverley Bell, Nigel Bell, and B. Cowie. *Formative assessment and science education*. Vol. 12. Springer Science & Business Media, 2001.

[23] Franklin Zaromb and Henry Roediger. "The testing effect in free recall is associated with enhanced organisation processes". In: *Memory & cognition* 38 (2010), pp. 995–1008. DOI: 10.3758/MC.38.8.995.

[24] H. E. Jones. "The effects of examination on the performance of learning". In: *Archives of Psychology* 10.1–70 (1923).

[25] Herbert F. Spitzer. "Studies in retention". In: *Journal of Educational Psychology* 30.9 (1939), p. 641.

[26] John A. Glover. "The testing phenomenon: Not gone but nearly forgotten". In: *Journal of Educational Psychology* 81.3 (1989), p. 392.

[27] Endel Tulving. "The effects of presentation and recall of material in free-recall learning". In: *Journal of verbal learning and verbal behavior* 6.2 (1967), pp. 175–184.

[28] Mark Carrier and Harold Pashler. "The influence of retrieval on retention". In: *Memory & Cognition* 20.6 (1992), pp. 633–642.

[29] C. Donald Morris, John Bransford, and Jeffery J. Franks. "Levels of processing versus transfer appropriate processing". In: *Journal of Verbal Learning and Verbal Behavior* 16 (1977), pp. 519–533. DOI: 10.1016/S0022-5371(77)80016-9.

[30] Henry Roediger, David A Gallo, and Lisa Geraci. "Processing approaches to cognition: The impetus from the levels-of-processing framework". In: *Memory (Hove, England)* 10 (2002), pp. 319–332. DOI: 10.1080/09658210224000144.

[31] Van Merrienboer, Jeroen J. G., Paul Kirschner, and P. Kester. "Taking the Load Off a Learner's Mind : Instructional Design for Complex Learning". In: *Educational Psychologist* 38 (2003). DOI: 10.1207/S15326985EP3801{\textunderscore}2.

[32] Lev Semenovich. Vygotsky. *Mind in society: The development of higher mental processes*. 1978.

[33] Carol S. Dweck. *Mindset: The new psychology of success*. Random House Digital, Inc, 2008.

[34] Carol S. Dweck. “Motivational processes affecting learning”. In: *American Psychologist* 41 (1986), pp. 1040–1048. DOI: 10.1037/0003-066X.41.10.1040.

[35] A. Elliot and C. S. Dweck. “Achievement motivation”. In: *Handbook of child psychology: social and personality development. New York: Wiley* (1983), pp. 643–691.

[36] Elaine S. Elliott and Carol S. Dweck. “Goals: An approach to motivation and achievement”. In: *Journal of Personality and Social Psychology* 54.1 (1988). DOI: 10.1037/0022-3514.54.1.5.

[37] John G. Nicholls. “Conceptions of ability and achievement motivation”. In: *Research on motivation in education* 1 (1984), pp. 39–73.

[38] C.-L.C. Kulik, J. A. Kulik, and R. L. Bangert-Drowns. “Effectiveness of Mastery Learning Programs: A Meta-Analysis”. In: *Review of Educational Research* 60 (1990), pp. 265–299. DOI: 10.3102/00346543060002265.

[39] James A Kulik, Chen-Lin Kulik, and Kevin Carmichael. “The Keller plan in science teaching”. In: *Science* 183.4123 (1974), pp. 379–383.

[40] James A. Kulik and Chen-Lin C. Kulik. “Effectiveness of the Personalized System of Instruction”. In: *Engineering Education* 66 (1975).

[41] James A. Kulik, Chen-Lin C. Kulik, and Peter A. Cohen. “A Meta-Analysis of Outcome Studies of Keller’s Personalized System of Instruction”. In: *American Psychologist* 34 (1979), pp. 307–318. DOI: 10.1037/0003-066X.34.4.307.

[42] Thomas Guskey and Sally L. Gates. “A Synthesis of Research on Group-Based Mastery Learning Programs”. In: (1985).

[43] Robert E Slavin. “Mastery learning reconsidered”. In: *Review of educational research* 57.2 (1987), pp. 175–213.

[44] Gloria Ladson-Billings. “From the Achievement Gap to the Education Debt: Understanding Achievement in U.S. Schools”. In: *Educational Researcher* 35.7 (2006), pp. 3–12.

[45] J. H. Block. *Schools, Society, and Mastery Learning*. Holt, Rinehart and Winston, 1974.

[46] Patrick Groff. “Some Criticisms of Mastery Learning”. In: *The Education Digest* 40.5 (1975), pp. 26–28.

[47] Zemira R. Mevarech and Shulamit Werner. “Are mastery learning strategies beneficial for developing problem solving skills?” In: *Higher Education* 14 (1985), pp. 425–432. DOI: 10.1007/BF00136514.

[48] Owen T. Anderson and Robert A. Artman. “A Self-Paced, Independent Study, Introductory Physics Sequence-Description and Evaluation”. In: *American Journal of Physics* 40 (1972), pp. 1737–1742. DOI: 10.1119/1.1987056.

[49] Michael A. Philippas and R. W. Sommerfeldt. “Keller vs Lecture Method in General Physics Instruction”. In: *American Journal of Physics* 40 (1972), pp. 1300–1306. DOI: 10.1119/1.1986818.

[50] Ben A. Green. “Physics teaching by the Keller Plan at MIT”. In: *American Journal of Physics* 39 (1971). DOI: 10.1119/1.1986280.

[51] G. C. Brown, K. I. Greisen, A. J. Sievers, and C. J. Naegele. “Self-paced introductory physics course-An eight-year progress report”. In: *American Journal of Physics* 45 (1977), pp. 1082–1088. DOI: 10.1119/1.10930.

[52] Graeme D. Putt. "Testing the mastery concept of self-paced learning in physics". In: *American Journal of Physics* 45 (1977), pp. 472–475. DOI: 10.1119/1.10824.

[53] John L. Safko. "Structure for a large enrollment, self-paced, mastery oriented astronomy course". In: *American Journal of Physics* 44 (1976), pp. 658–662. DOI: 10.1119/1.10330.

[54] Gerd Kortemeyer, E. Kashy, W. Benenson, and Wolfgang Bauer. "Experiences using the open-source learning content management and assessment system LON-CAPA in introductory physics courses". In: *American Journal of Physics* 76 (2008), pp. 438–444. DOI: 10.1119/1.2835046.

[55] Elsa-Sofia Morote and David Pritchard. "Technology Closes the gap between Students' Individual Skills and Background Differences". In: *Proceedings of Society for Information Technology & Teacher Education International Conference 2004*. Ed. by Richard Ferdig, Caroline Crawford, Roger Carlsen, Niki Davis, Jerry Price, Roberta Weber, and Dee Anna Willis. Atlanta, GA, USA: Association for the Advancement of Computing in Education (AACE), 2004, pp. 826–831.

[56] Kurt Vanlehn, Collin Lynch, Kay G. Schulze, Joel Shapiro, Robert Shelby, Linwood Taylor, Donald Treacy, Anders Weinstein, and Mary Wintersgill. "The Andes Physics Tutoring System: Lessons Learned". In: *I. J. Artificial Intelligence in Education* 15 (2005), pp. 147–204.

[57] Raymond W. Kulhavy and Walter Wager. "Feedback in programmed instruction: historical context and implications for practice". In: *Interactive instruction and feedback*. Ed. by John V. Dempsey and Gregory C. Sales. Englewood Cliffs, NJ: Educational Technology Publications, 1993.

[58] John Hattie and Helen Timperley. "The Power of Feedback". In: *Review of Educational Research* 77.1 (2007), pp. 81–112. DOI: 10.3102/003465430298487.

[59] Fabienne M. Van der Kleij, Remco C. W. Feskens, and Theo J. H. M. Eggen. "Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis". In: *Review of Educational Research* 85.4 (2015), pp. 475–511. DOI: 10.3102/0034654314564881.

[60] Pearson. *Results Library*.

[61] Brendon D. Mikula and Andrew Heckler. "Framework and implementation for improving physics essential skills via computer-based practice: Vector math". In: *Physical Review Physics Education Research* 13 (2017). DOI: 10.1103/PhysRevPhysEducRes.13.010122.

[62] Andrew Heckler and Brendon D. Mikula. "Factors affecting learning of vector math from computer-based practice: Feedback complexity and prior knowledge". In: *Physical Review Physics Education Research* 12 (2016). DOI: 10.1103/PhysRevPhysEducRes.12.010134.

[63] Ward Canfield. "ALEKS: a Web-based intelligent tutoring system". In: *Mathematics and Computer Education* 35 (2001), pp. 152–158.

[64] Noah Schroeder, Gary Gladding, Brianne Gutmann, and Timothy Stelzer. "Narrated animated solution videos in a mastery setting". In: *Physical Review Special Topics-Physics Education Research* 11.1 (2015), p. 010103.

[65] Noah Schroeder. "Mastery-style exercises in physics". PhD thesis. Urbana, IL: University of Illinois at Urbana-Champaign, 2015.

[66] Richard Mayer. "The Cambridge Handbook Of Multimedia Learning". In: *The Cambridge Handbook of Multimedia Learning* (2005). DOI: 10.1017/CBO9780511816819.

[67] Wolfgang Schnotz. "Commentary: Towards an Integrated View of Learning from Text and Visual Displays". In: *Educational Psychology Review* 14.1 (2002), pp. 101–120. DOI: 10.1023/A:1013136727916.

[68] Richard E. Mayer, Julie Heiser, and Steve Lonn. "Cognitive constraints on multimedia learning: When presenting more material results in less understanding". In: *Journal of educational psychology* 93.1 (2001), p. 187.

[69] Slava Kalyuga, Paul Chandler, and John Sweller. "When Redundant On-Screen Text in Multimedia Technical Instruction Can Interfere With Learning". In: *Human factors* 46 (2004), pp. 567–581. DOI: 10.1518/hfes.46.3.567.1640.

[70] John Sweller. "The worked example effect and human cognition". In: *Learning and Instruction - LEARN INSTR* 16 (2006), pp. 165–169. DOI: 10.1016/j.learninstruc.2006.02.005.

[71] Slava Kalyuga, Paul Ayres, Paul Chandler, and John Sweller. "The Expertise Reversal Effect". In: *Educational Psychologist* 38.1 (2003), pp. 23–31. DOI: 10.1207/S15326985EP3801/_4.

[72] Jana Reisslein, Robert K. Atkinson, Patrick Seeling, and Martin Reisslein. "Encountering the Expertise Reversal Effect with a Computer-based Environment on Electrical Circuit Analysis". In: *Learning and Instruction* 16 (2006), pp. 92–103. DOI: 10.1016/j.learninstruc.2006.02.008.

[73] Kalyuga Slava. "Expertise Reversal Effect and Its Implications for Learner-Tailored Instruction". In: *Educational Psychology Review* 4 (2007), p. 509.

[74] Nicholas J. Cepeda, Harold Pashler, Edward Vul, John T. Wixted, and Doug Rohrer. "Distributed practice in verbal recall tasks: A review and quantitative synthesis". In: *Psychological bulletin* 132.3 (2006), p. 354.

[75] Vanessa D. Moss. "The Efficacy of Massed Versus Distributed Practice As a Function of Desired Learning Outcomes and Grade Level of the Student". PhD thesis. Utah State University, 1995.

[76] Michael G. Grote. "Distributed versus massed practice in high school physics". In: *School Science and Mathematics* 95.2 (1995), pp. 97–101.

[77] John Risley. *WebAssign*. 1997.

[78] Cengage. *Enhanced WebAssign Case Studies*.

[79] Pearson. *Mastering Physics*.

[80] Dale Dawes. "Effectively Implementing an Online Homework and Testing Management System to Increase Student Achievement - A Student Tailored Pedagogical Approach". PhD thesis. Columbia University, 2016.

[81] Tolga Gok. "Using Computer-Assisted Personalized Assignment System in a Large-Enrollment General Physics". In: *European Journal of Physics Education* 1.1 (2010), pp. 28–43.

[82] E. Kashy, B. M. Sherrill, Y. Tsai, D. Thaler, D. Weinshank, M. Engelmann, and D. J. Morrissey. "CAPA—An integrated computer-assisted personalized assignment system". In: *American Journal of Physics* 61.12 (1993), pp. 1124–1130.

[83] E. Kashy, S. J. Gaff, N. H. Pawley, W. L. Stretch, S. L. Wolfe, D. J. Morrissey, and Y. Tsai. "Conceptual questions in computer-assisted assignments". In: *American Journal of Physics* 63.11 (1995), pp. 1000–1005.

[84] Gerd Kortemeyer. "An analysis of asynchronous online homework discussions in introductory physics courses". In: *American Journal of Physics* 74.6 (2006), pp. 526–536.

[85] M. Thoennessen and M. J. Harrison. "Computer-assisted assignments in a large physics class". In: *Computers & Education* 27.2 (1996), pp. 141–147.

[86] Jose Mestre, David M. Hart, Kenneth A. Rath, and Robert Dufresne. "The Effect of Web-Based Homework on Test Performance in Large Enrollment Introductory Physics Courses". In: *Journal of Computers in Mathematics and Science Teaching* 21.3 (2002), pp. 229–251.

[87] Andrea M. Pascarella. "The influence of web-based homework on quantitative problem-solving in a university physics class". In: *Proc. NARST Annual Meeting*. 2004.

[88] Scott Bonham, Robert Beichner, and Duane Deardorff. "Online homework: Does it make a difference?" In: *The Physics Teacher* 39.5 (2001), pp. 293–296.

[89] Scott W. Bonham, Duane L. Deardorff, and Robert J. Beichner. "Comparison of student performance using web and paper-based homework in college-level physics". In: *Journal of research in science teaching* 40.10 (2003), pp. 1050–1071.

[90] Alan C. Bugbee Jr. "The equivalence of paper-and-pencil and computer-based testing". In: *Journal of research on computing in education* 28.3 (1996), pp. 282–299.

[91] W. R. Schleter and R. M. Bennett. "Using web-based homework in an introductory engineering physics course". In: *Proceedings, ASEE Annual Convention, Paper*. Vol. 2279. 2006.

[92] L. Hassler, L. Dennis, H. Ng, C. Johnson, D. Ossont, G. Ogawa, and C. Nahmias. "Computer-assisted vs. traditional homework: results of a pilot research project". In: *WIT Transactions on Information and Communication Technologies* 31 (2004).

[93] K. Kelvin Cheng, Beth Ann Thacker, Richard L. Cardenas, and Catherine Crouch. "Using an online homework system enhances students' learning of physics concepts in an introductory physics course". In: *American Journal of Physics* 72.11 (2004), pp. 1447–1453.

[94] Alan C. Bugbee Jr and Frank M. Bernt. "TIME'S UP! The Effects of Time Constraints and Mode of Administration on Test Performance and Perception". In: *Journal of Research on Computing in Education* 25.2 (1992), pp. 243–253.

[95] Guoqing Tang and Aaron Titus. "Increasing students' time on task in calculus and general physics courses through WebAssign". In: *age* 7 (2002), p. 1.

[96] Arthur W. Chickering and Zelda F. Gamson. "Seven principles for good practice in undergraduate education". In: *AAHE bulletin* 3 (1987), p. 7.

[97] John R. Anderson, Albert T. Corbett, Kenneth R. Koedinger, and Ray Pelletier. "Cognitive tutors: Lessons learned". In: *The journal of the learning sciences* 4.2 (1995), pp. 167–207.

[98] Andrew C. Butler, Jeffrey D. Karpicke, and Henry L. Roediger III. "The effect of type and timing of feedback on learning from multiple-choice tests". In: *Journal of Experimental Psychology: Applied* 13.4 (2007), p. 273.

[99] Valerie J. Shute. "Focus on formative feedback". In: *Review of educational research* 78.1 (2008), pp. 153–189.

[100] Andrew C. Butler and Henry L. Roediger. "Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing". In: *Memory & Cognition* 36.3 (2008), pp. 604–616.

[101] D. Kane and B. Sherwood. "A Computer-based Course in Classical Mechanics". In: *Computers and Education* 4.1 (1980), pp. 15–36.

[102] L. M. Jones, Dennis Kane, Bruce Arne Sherwood, and R. A. Avner. "A final-exam comparison involving computer-based instruction". In: *American Journal of Physics* 51.6 (1983), pp. 533–538.

[103] D. M. Raineri, B. G. Mehrtens, and A. W. Hubler. "CyberProf^{TM} - An intelligent human-computer interface for interactive instruction on the world wide web". In: *Journal of Asynchronous Learning Network* 1.2 (1997), pp. 20–36.

[104] Alfred W. Hübler and Andrew M. Assad. "CyberProf: An Intelligent Human-Computer Interface for Asynchronous Wide Area Training and Teaching". In: *World Wide Web Journal* 1 (1996).

[105] Tim Stelzer and Gary Gladding. "The evolution of Web-based activities in physics at Illinois". In: *Newsletter of the Forum on Education of the American Physical Society, Fall.* 2001, pp. 7–8.

[106] Slava Kalyuga, Paul Ayres, Paul Chandler, and John Sweller. "The expertise reversal effect". In: *Educational Psychologist* 38.1 (2003), pp. 23–31.

[107] Gary Gladding, Brianne Gutmann, Noah Schroeder, and Timothy Stelzer. "Clinical study of student learning using mastery style versus immediate feedback online activities". In: *Physical Review Special Topics-Physics Education Research* 11.1 (2015), p. 010114.

[108] Michelene T.H. Chi, Paul J. Feltovich, and Robert Glaser. "Categorization and representation of physics problems by experts and novices". In: *Cognitive Science* 5.2 (1981), pp. 121–152.

[109] Angela T. Barlow, Natasha E. Gerstenschlager, Jeremy F. Strayer, Alyson E. Lischka, D. Christopher Stephens, Kristin S. Hartland, and J. Christopher Willingham. "Scaffolding for Access to Productive Struggle". In: *Mathematics Teaching in the Middle School* 23.4 (2018), pp. 202–207.

[110] Carina Granberg. "Discovering and addressing errors during mathematics problem-solving—A productive struggle?" In: *The Journal of Mathematical Behavior* 42 (2016), pp. 33–48.

[111] Sharyn Livy, Tracey Muir, Peter Sullivan, et al. "Challenging tasks lead to productive struggle!" In: *Australian Primary Mathematics Classroom* 23.1 (2018), p. 19.

[112] Sararose D. Lynch, Jessica H. Hunt, and Katherine E. Lewis. "Productive Struggle for All: Differentiated Instruction". In: *Mathematics Teaching in the Middle School* 23.4 (2018), pp. 194–201.

[113] Jaclyn M. Murawska. "Seven Billion People: Fostering Productive Struggle". In: *Mathematics Teaching in the Middle School* 23.4 (2018), pp. 208–214.

[114] Marian Pasquale. "Productive Struggle in Mathematics. Interactive STEM Research+ Practice Brief". In: *Education Development Center, Inc* (2016).

[115] Cynthia Townsend, David Slavit, and Amy Roth McDuffie. "Supporting All Learners in Productive Struggle". In: *Mathematics Teaching in the Middle School* 23.4 (2018), pp. 216–224.

[116] Hiroko K. Warshauer. "Strategies to support productive struggle". In: *Mathematics Teaching in the Middle School* 20.7 (2015), pp. 390–393.

[117] Hiroko Kawaguchi Warshauer. "Productive struggle in middle school mathematics classrooms". In: *Journal of Mathematics Teacher Education* 18.4 (2015), pp. 375–400.

[118] Manu Kapur. "Productive Failure in Learning Math". In: *Cognitive Science* 38.5 (2014), pp. 1008–1022. DOI: 10.1111/cogs.12107.

[119] Manu Kapur. "A further study of productive failure in mathematical problem solving: unpacking the design components". In: *Instructional Science* 39.4 (2011), pp. 561–579. DOI: 10.1007/s11251-010-9144-3.

[120] Manu Kapur. "Learning from productive failure". In: *Learning: Research and Practice* 1.1 (2015), pp. 51–65. DOI: 10.1080/23735082.2015.1002195.

[121] Manu Kapur and Katerine Bielaczyc. "Designing for Productive Failure". In: *Journal of the Learning Sciences* 21.1 (2012), pp. 45–83. DOI: 10.1080/10508406.2011.591717.

[122] Inga Glogger, Julian Kappich, Rolf Schwonke, Lars Holzapfel, Matthias Nuckles, and Alexander Renkl. "Inventing prepares learning motivationally, but a worked-out solution enhances learning outcomes". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 35. 2013.

[123] Katharina Loibl and Nikol Rummel. "Knowing what you don't know makes failure productive". In: *Learning and Instruction* 34 (2014), pp. 74–85. DOI: 10.1016/j.learninstruc.2014.08.004.

[124] Carol S. Dweck. "Mindsets and math/science achievement". In: (2014).

[125] Scott W. Mcquiggan, Sunyoung Lee, and James C. Lester. "Early prediction of student frustration". In: *International Conference on Affective Computing and Intelligent Interaction*. 2007, pp. 698–709.

[126] Marcos D. Caballero, Bethany R. Wilcox, Rachel E. Pepper, and Steven J. Pollock. "ACER: a framework on the use of mathematics in upper-division physics". In: *AIP Conference Proceedings*. Vol. 1513. 2013, pp. 90–93.

[127] Bethany R. Wilcox, Marcos D. Caballero, Daniel A. Rehn, and Steven J. Pollock. "Analytic framework for students' use of mathematics in upper-division physics". In: *Physical Review Special Topics-Physics Education Research* 9.2 (2013), p. 020119.

[128] Dimitri R. Dounas-Frazer, Jacob T. Stanley, and H. J. Lewandowski. "Student ownership of projects in an upper-division optics laboratory course: A multiple case study of successful experiences". In: *Phys. Rev. Phys. Educ. Res.* 13.2 (2017), p. 020136. DOI: 10.1103/PhysRevPhysEducRes.13.020136.

[129] Noah Podolefsky. "Intentional design for empowerment". In: *arXiv preprint arXiv:1308.5956* (2013).

[130] Jean Lave and Ray McDermott. "Estranged labor learning". In: *Outlines. Critical Practice Studies* 4.1 (2002), pp. 19–48.

[131] Angela Little. "Proudness: What is it? why is it important? and how do we design for it in college physics and astronomy education". In: *STATUS: A Report on Women in Astronomy, Newsletter published by the American Astronomical Society* (2015).

[132] David I. Hanauer, Mark J. Graham, and Graham F. Hatfull. "A Measure of College Student Persistence in the Sciences (PITS)". In: *CBE—Life Sciences Education* 15.4 (2016), ar54. DOI: 10.1187/cbe.15-09-0185.

[133] Marina Milner-Bolotin. "The effects of topic choice in project-based instruction on undergraduate physical science students' interest, ownership, and motivation". Ph.D. Austin, TX: University of Texas at Austin, 2001.

[134] Margareta Enghag. "Two dimensions of student ownership of learning during small-group work with miniprojects and context rich problems in physics". PhD thesis. Mälardalen University, Eskilstuna.

[135] Candice R. Stefanou, Kathleen C. Perencevich, Matthew DiCintio, and Julianne C. Turner. "Supporting Autonomy in the Classroom: Ways Teachers Encourage Student Decision Making and Ownership". In: *Educational Psychologist* 39.2 (2004), pp. 97–110. DOI: 10.1207/s15326985ep3902/_2.

[136] Brianne Gutmann. "Mastery learning in the zone of proximal development". In: *Physics Education Research Conference 2017*. PER Conference. Cincinnati, OH, 2017, pp. 156–159.

[137] Eugene Torigoe and Gary Gladding. "Same to Us, Different to Them: Numeric Computation versus Symbolic Representation". In: *AIP Conference Proceedings* 883.1 (2007), pp. 153–156. DOI: 10.1063/1.2508715.

[138] Eugene Torigoe and Gary Gladding. "Symbols: Weapons of Math Destruction". In: *AIP Conference Proceedings* 951.1 (2007), pp. 200–203. DOI: 10.1063/1.2820933.

[139] Penny L. Beed, E. Marie Hawkins, and Cathy M. Roller. "Moving learners toward independence: The power of scaffolded instruction". In: *The Reading Teacher* 44.9 (1991), pp. 648–655.

[140] Barak Rosenshine and Carla Meister. "The use of scaffolds for teaching higher-level cognitive strategies". In: *Educational leadership* 49.7 (1992), pp. 26–33.

[141] David Wood, Jerome S. Bruner, and Gail Ross. "The role of tutoring in problem solving". In: *Journal of child psychology and psychiatry* 17.2 (1976), pp. 89–100.

[142] Katherine F. Schetz and Andrew J. Stremmel. "Teacher-assisted computer implementation: A Vygotskian perspective". In: *Early Education and Development* 5.1 (1994), pp. 18–26.

[143] Brianne Gutmann. "Effective Grain-Size of Mastery-Style Online Homework Levels". In: *Physics Education Research Conference 2018*. PER Conference. Washington, DC, 2018.

[144] Avraham N. Kluger and Angelo DeNisi. "The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory". In: *Psychological Bulletin* 119.2 (1996). DOI: 10.1037/0033-2909.119.2.254.

[145] Richard S Lysakowski and Herbert J Walberg. "Instructional Effects of Cues, Participation, and Corrective Feedback: A Quantitative Synthesis". In: *American Educational Research Journal* 19.4 (1982), pp. 559–572. DOI: 10.3102/00028312019004559.

[146] Bangert-Drowns Robert L., Kulik James A., and Kulik Chen-Lin C. "Effects of Frequent Classroom Testing". In: *The Journal of Educational Research* 85.2 (1991), p. 89.

[147] James A. Kulik and Chen-Lin C. Kulik. "Timing of Feedback and Verbal Learning". In: *Review of Educational Research* 58.1 (1988), pp. 79–97. DOI: 10.3102/00346543058001079.

[148] Gülşah Başol and George Johanson. "Effectiveness of frequent testing over achievement: A meta analysis study". In: *Journal of Human Sciences* 6.2 (2009).

[149] John Hattie. "Influences on student learning". In: *Inaugural lecture given on August* 2 (1999), p. 1999.

[150] George Mandler and Seymour B. Sarason. *A study of anxiety and learning*. US, 1952. DOI: 10.1037/h0062855.

[151] Jerrell C. Cassady and Ronald E. Johnson. "Cognitive Test Anxiety and Academic Performance". In: *Contemporary Educational Psychology* 27.2 (2002), pp. 270–295. DOI: 10.1006/ceps.2001.1094.

[152] Ray Hembree. "Correlates, Causes, Effects, and Treatment of Test Anxiety". In: *Review of Educational Research* 58.1 (1988), pp. 47–77. DOI: 10.3102/00346543058001047.

[153] Frank E. Fulkerson and Glen Martin. "Effects of Exam Frequency on Student Performance, Evaluations of Instructor, and Test Anxiety". In: *Teaching of Psychology* 8.2 (1981), pp. 90–93. DOI: 10.1207/s15328023top0802/_7.

[154] Erkan Sulun, Ertem Nalbantoglu, and Emine Kivanc Oztug. "The effect of exam frequency on academic success of undergraduate music students and comparison of students performance anxiety levels". In: *Quality & Quantity* 52.1 (2018), pp. 737–752.

[155] Austin H. Turney. "The effect of frequent short objective tests upon the achievement of college students in educational psychology". In: *School and Society* 33.858 (1931), pp. 760–762.

[156] Daniel H. Kulp. "Weekly tests for graduate students". In: *School and Society* 38.970 (1933), pp. 157–159.

[157] V. T. Mawhinney, D. E. Bostow, D. R. Laws, G. J. Blumenfeld, and B. L. Hopkins. "A comparison of students studying – Behavior produced by daily, weekly, and three-week testing schedules". In: *Journal of Applied Behavior Analysis* 4.4 (1971), pp. 257–264.

[158] Terence Nip, Elsa L. Gunter, Geoffrey L. Herman, Jason W. Morphew, and Matthew West. "Using a computer-based testing facility to improve student learning in a programming languages and compilers course". In: *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. 2018, pp. 568–573.

[159] Binglin Chen, Matthew West, and Craig Zilles. "How much randomization is needed to deter collaborative cheating on asynchronous exams?" In: *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. 2018, p. 62.

[160] Ward Mitchell Cates. "The efficacy of retesting in relation to improved test performance of college undergraduates". In: *The Journal of Educational Research* 75.4 (1982), pp. 230–236.

[161] Herbert Friedman. "Repeat examinations in introductory statistics courses". In: *Teaching of Psychology* 14.1 (1987), pp. 20–23.

[162] Allison M. Geving, Shannon Webb, and Bruce Davis. "Opportunities for repeat testing: Practice doesn't always make perfect". In: *Applied HRM Research* 10.2 (2005), pp. 47–56.

[163] James A. Kulik, Chen-Lin C. Kulik, and Robert L. Bangert. "Effects of Practice on Aptitude and Achievement Test Scores". In: *American Educational Research Journal* 21.2 (1984), pp. 435–447.

[164] Sandra M. Juhler, Janice F. Rech, Steven G. From, and Monica M. Brogan. "The effect of optional retesting on college students' achievement in an individualized algebra course". In: *The Journal of experimental education* 66.2 (1998), pp. 125–137.

[165] Michelene T. H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. "Self-explanations: How students study and use examples in learning to solve problems". In: *Cognitive science* 13.2 (1989), pp. 145–182.

[166] Marc Schwartz. "Khan academy: The illusion of understanding". In: *Online Learning Journal* 17.4 (2013).

[167] Jörg Wittwer and Alexander Renkl. "Why Instructional Explanations Often Do Not Work: A Framework for Understanding the Effectiveness of Instructional Explanations". In: *Educational Psychologist* 43.1 (2008), pp. 49–64. DOI: 10.1080/00461520701756420.

[168] Michelene T. H. Chi, Nicholas de Leeuw, Mei-hung Chiu, and Christian LaVancher. "Eliciting Self-Explanations Improves Understanding". In: *Cognitive Science: A Multidisciplinary Journal of Artificial Intelligence, Linguistics, Neuroscience, Philosophy, Psych* 18.3 (1994), pp. 439–477.

[169] Jayson M. Nissen and Jonathan T. Shemwell. "Gender, experience, and self-efficacy in introductory physics". In: *Phys. Rev. Phys. Educ. Res.* 12.2 (2016), p. 020105. DOI: 10.1103/PhysRevPhysEducRes.12.020105.

[170] Maria Ong, Carol Wright, Lorelle L. Espinosa, and Gary Orfield. "Inside the Double Bind: A Synthesis of Empirical Research on Undergraduate and Graduate Women of Color in Science, Technology, Engineering,and Mathematics". In: *Harvard Educational Review* 81.2 (2011), pp. 172–209.

[171] Vashti Sawtelle, Eric Brewe, and Laird H. Kramer. "Exploring the relationship between self-efficacy and retention in introductory physics". In: *Journal of Research in Science Teaching* 49.9 (2012), pp. 1096–1121. DOI: 10.1002/tea.21050.

[172] Adrienne Traxler and Eric Brewe. "Equity investigation of attitudinal shifts in introductory physics". In: *Phys. Rev. ST Phys. Educ. Res.* 11.2 (2015), p. 020132. DOI: 10.1103/PhysRevSTPER.11.020132.

[173] Emily M. Marshman, Z. Yasemin Kalender, Timothy Nokes-Malach, Christian Schunn, and Chandralekha Singh. "Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm?" In: *Phys. Rev. Phys. Educ. Res.* 14.2 (2018), p. 020123. DOI: 10.1103/PhysRevPhysEducRes.14.020123.

[174] Karyn L. Lewis, Jane G. Stout, Steven J. Pollock, Noah D. Finkelstein, and Tiffany A. Ito. "Fitting in or opting out: A review of key social-psychological factors influencing a sense of belonging for women in physics". In: *Phys. Rev. Phys. Educ. Res.* 12.2 (2016), p. 020110. DOI: 10.1103/PhysRevPhysEducRes.12.020110.

[175] Sarah L. Eddy and Sara E. Brownell. "Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines". In: *Phys. Rev. Phys. Educ. Res.* 12.2 (2016), p. 020106. DOI: 10.1103/PhysRevPhysEducRes.12.020106.

[176] Adrienne L. Traxler, Ximena C. Cid, Jennifer Blue, and Ramón Barthelemy. "Enriching gender in physics education research: A binary past and a complex future". In: *Phys. Rev. Phys. Educ. Res.* 12.2 (2016), p. 020114. DOI: 10.1103/PhysRevPhysEducRes.12.020114.

[177] Megan Bruun, Shannon Willoughby, and Jessi L. Smith. "Identifying the stereotypical who, what, and why of physics and biology". In: *Phys. Rev. Phys. Educ. Res.* 14.2 (2018), p. 020125. DOI: 10.1103/PhysRevPhysEducRes.14.020125.

[178] Alexandru Maries, Nafis I. Karim, and Chandralekha Singh. "Is agreeing with a gender stereotype correlated with the performance of female students in introductory physics?" In: *Phys. Rev. Phys. Educ. Res.* 14.2 (2018), p. 020119. DOI: 10.1103/PhysRevPhysEducRes.14.020119.

[179] Zahra Hazari, Philip M. Sadler, and Gerhard Sonnert. "The science identity of college students: Exploring the intersection of gender, race, and ethnicity". In: *Journal of College Science Teaching* 42.5 (2013), pp. 82–91.

[180] Lauren E. Kost-Smith, Steven J. Pollock, and Noah D. Finkelstein. "Gender disparities in second-semester college physics: The incremental effects of a "smog of bias"". In: *Phys. Rev. ST Phys. Educ. Res.* 6.2 (2010), p. 020112. DOI: 10.1103/PhysRevSTPER.6.020112.

[181] Seyranian, Viviane and Madva, Alex and Duong, Nicole and Abramzon, Nina and Tibbetts, Yoi and Harackiewicz, Judith M. "The longitudinal effects of STEM identity and gender on flourishing and achievement in college physics". In: *International Journal of STEM Education* 5.1 (2018), p. 40. DOI: 10.1186/s40594-018-0137-0.

[182] Katemari Rosa and Felicia Moore Mensah. "Educational pathways of Black women physicists: Stories of experiencing and overcoming obstacles in life". In: *Phys. Rev. Phys. Educ. Res.* 12.2 (2016), p. 020113. DOI: 10.1103/PhysRevPhysEducRes.12.020113.

[183] Angela M. Kelly. "Physics teachers' perspectives on factors that affect urban physics participation and accessibility". In: *Phys. Rev. ST Phys. Educ. Res.* 9.1 (2013), p. 010122. DOI: 10.1103/PhysRevSTPER.9.010122.

[184] N. Reid and E. A. Skryabina. "Gender and physics". In: *International Journal of Science Education* 25.4 (2003), pp. 509–536.

[185] Lauren E. Kost, Steven J. Pollock, and Noah D. Finkelstein. "Characterizing the gender gap in introductory physics". In: *Phys. Rev. ST Phys. Educ. Res.* 5.1 (2009), p. 010101. DOI: 10.1103/PhysRevSTPER.5.010101.

[186] Idaykis Rodriguez, Geoff Potvin, and Laird H. Kramer. "How gender and reformed introductory physics impacts student success in advanced physics courses and continuation in the physics major". In: *Phys. Rev. Phys. Educ. Res.* 12.2 (2016), p. 020118. DOI: 10.1103/PhysRevPhysEducRes.12.020118.

[187] Paul W. Irving and Eleanor C. Sayre. "Becoming a physicist: The roles of research, mindsets, and milestones in upper-division student perceptions". In: *Phys. Rev. ST Phys. Educ. Res.* 11.2 (2015), p. 020120. DOI: 10.1103/PhysRevSTPER.11.020120.

[188] Vashti Sawtelle, Eric Brewe, Renee Michelle Goertzen, and Laird H. Kramer. "Identifying events that impact self-efficacy in physics learning". In: *Phys. Rev. ST Phys. Educ. Res.* 8.2 (2012), p. 020111. DOI: 10.1103/PhysRevSTPER.8.020111.

[189] Eleanor W. Close, Jessica Conn, and Hunter G. Close. "Becoming physics people: Development of integrated physics identity through the Learning Assistant experience". In: *Phys. Rev. Phys. Educ. Res.* 12.1 (2016), p. 010109. DOI: 10.1103/PhysRevPhysEducRes.12.010109.

[190] Robynne M. Lock, Zahra Hazari, and Geoff Potvin. "Physics career intentions: The effect of physics identity, math identity, and gender". In: *AIP Conference Proceedings* 1513.1 (2013), pp. 262–265. DOI: 10.1063/1.4789702.

[191] Zahra Hazari, Gerhard Sonnert, Philip M. Sadler, and Marie-Claire Shanahan. "Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study". In: *Journal of Research in Science Teaching* 47.8 (2010), pp. 978–1003. DOI: 10.1002/tea.20363.

[192] Akira Miyake, Lauren E. Kost-Smith, Noah D. Finkelstein, Steven J. Pollock, Geoffrey L. Cohen, and Tiffany A. Ito. "Reducing the Gender Achievement Gap in College Science: A Classroom Study of Values Affirmation". In: *Science* 330.6008 (2010), pp. 1234–1237. DOI: 10.1126/science.1195996.

[193] Staffan Andersson and Anders Johansson. "Gender gap or program gap? Students' negotiations of study practice in a course in electromagnetism". In: *Phys. Rev. Phys. Educ. Res.* 12.2 (2016), p. 020112. DOI: 10.1103/PhysRevPhysEducRes.12.020112.

[194] Rochelle Gutierrez. "A Gap-Gazing Fetish in Mathematics Education? Problematizing Research on the Achievement Gap". In: *Journal for Research in Mathematics Education* 39.4 (2008), pp. 357–364.

[195] Anna Danielsson. "Gender in physics education research: A review and a look forward". In: (2010).

[196] Kimberle Crenshaw. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics". In: *University of Chicago Legal Forum*. Vol. 1989. 1989, p. 8.

[197] Sarah Theule Lubienski. "On gap gazing in mathematics education: The need for gaps analyses". In: *Journal for Research in Mathematics Education* (2008), pp. 350–356.

[198] Claude M. Steele, Steven J. Spencer, and Joshua Aronson. "Contending with group image: The psychology of stereotype and social identity threat". In: vol. 34. Advances in Experimental Social Psychology. Academic Press, 2002, pp. 379–440. DOI: 10.1016/S0065-2601(02)80009-0.

[199] Gregory M. Walton and Steven J. Spencer. "Latent Ability: Grades and Test Scores Systematically Underestimate the Intellectual Ability of Negatively Stereotyped Students". In: *Psychological Science* 20.9 (2009), p. 1132.

[200] Geoffrey L. Cohen and David K. Sherman. *The Psychology of Change: Self-Affirmation and Social Psychological Intervention*. 2014. DOI: 10.1146/annurev-psych-010213-115137.

[201] Adrian Madsen, Sarah B. McKagan, and Eleanor C. Sayre. *Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?* 2013. DOI: 10.1103/PhysRevSTPER.9.020121.

[202] D. S. Yeager and G. M. Walton. "Social-psychological interventions in education: They're not magic". In: *Review of Educational Research* 81.2 (2011), pp. 267–301. DOI: 10.3102/0034654311405999.

[203] Nurit Shnabel, Valerie Purdie-Vaughns, Jonathan E. Cook, Julio Garcia, and Geoffrey L. Cohen. *Demystifying Values-Affirmation Interventions: Writing About Social Belonging Is a Key to Buffering Against Identity Threat*. 2013. DOI: 10.1177/0146167213480816.

[204] Steve Stoessner and Catherine Good. *Stereotype Threat: An Overview*. 2011.

[205] Claude M. Steele and Joshua Aronson. "Stereotype Threat and the Intellectual Test Performance of African Americans". In: *Journal of Personality and Social Psychology* 69.5 (1995), pp. 797–811. DOI: 10.1037/0022-3514.69.5.797.

[206] Steven J. Spencer, Claude M. Steele, and Diane M. Quinn. "Stereotype threat and women's math performance". In: *Journal of experimental social psychology* 35.1 (1999), pp. 4–28.

[207] Margaret Walsh, Crystal Hickey, and Jim Duffy. "Influence of item content and stereotype situation on gender differences in mathematical problem solving". In: *Sex Roles* 41.3-4 (1999), pp. 219–240.

[208] Catherine Good, Joshua Aronson, and Jayne Ann Harder. "Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses". In: *Journal of applied developmental psychology* 29.1 (2008), pp. 17–28.

[209] M. Inzlicht and T. Ben-Zeev. "A Threatening Intellectual Environment: Why Females Are Susceptible to Experiencing Problem-Solving Deficits in the Presence of Males". In: *Psychological Science* 11.5 (2000), pp. 365–371.

[210] N. Ambady, M. Shih, A. Kim, and T. L. Pittinsky. "Stereotype susceptibility in children: Effects of Identity Activation on Quantitative Performance". In: *Psychological Science* 12.5 (2001), pp. 385–390. DOI: 10.1111/1467-9280.00371.

[211] Jean-Claude Croizet and Theresa Claire. "Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds". In: *Personality and Social Psychology Bulletin* 24.6 (1998), pp. 588–594.

[212] Lisa A. Harrison, Chiesha M. Stevens, Adrienne N. Monty, and Christine A. Coakley. "The consequences of stereotype threat on the academic performance of White and non-White lower income college students". In: *Social Psychology of Education* 9.3 (2006), pp. 341–357.

[213] D. K. Sherman, K. A. Hartson, K. R. Binning, V. Purdie-Vaughns, J. Garcia, S. Taborsky-Barba, S. Tomassetti, A. David Nussbaum, and G. L. Cohen. "Deflecting the trajectory and changing the narrative: How self-affirmation affects academic performance and motivation under identity threat". In: *Journal of Personality and Social Psychology* 104.4 (2013), pp. 591–618.

[214] Patricia M. Gonzales, Hart Blanton, and Kevin J. Williams. "The effects of stereotype threat and double-minority status on the test performance of Latino women". In: *Personality and Social Psychology Bulletin* 28.5 (2002), pp. 659–670.

[215] Toni Schmader and Michael Johns. "Converging evidence that stereotype threat reduces working memory capacity". In: *Journal of Personality and Social Psychology* 85.3 (2003), p. 440.

[216] B. Levy. "Improving memory in old age through implicit self-stereotyping". In: *Journal of Personality and Social Psychology* 71.6 (1996), pp. 1092–1107.

[217] J. K. Bosson, E. L. Haymovitz, and E. C. Pinel. "When saying and doing diverge: The effects of stereotype threat on self-reported versus non-verbal anxiety". In: *Journal of experimental social psychology* 40.2 (2004). DOI: 10.1016/S0022-1031(03)00099-4.

[218] Jeff Stone, Mike Sjomeling, Christian I. Lynch, and John M. Darley. "Stereotype Threat Effects on Black and White Athletic Performance". In: *Journal of Personality and Social Psychology* 77.6 (1999), pp. 1213–1227. DOI: 10.1037/0022-3514.77.6.1213.

[219] L. J. Kray, A. D. Galinsky, and L. Thompson. "Reversing the gender gap in negotiations: An exploration of stereotype regeneration". In: *Organizational Behavior and Human Decision Processes* 87.2 (2002). DOI: 10.1006/obhd.2001.2979.

[220] Yeung, Nai, Chi, Jonathan, and von Hippel, Courtney. "Stereotype threat increases the likelihood that female drivers in a simulator run over jaywalkers". In: *Accident Analysis and Prevention* 40.2 (2008). DOI: 10.1016/j.aap.2007.09.003.

[221] C. M. Steele, S. J. Spencer, and J. Aronson. *Contending with group image: The psychology of stereotype and social identity threat: Contending with group image: The psychology of stereotype and social identity threat.* Vol. 34. Advances in Experimental Social Psychology. 2002.

[222] C. M. Steele. "A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance: A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance". In: *American Psychologist* 52.6 (1997), pp. 613–629. DOI: 10.1037/0003-066X.52.6.613.

[223] Anne C. Krendl, Jennifer A. Richeson, William M. Kelley, and Todd F. Heatherton. "The Negative Consequences of Threat: A Functional Magnetic Resonance Imaging Investigation of the Neural Mechanisms Underlying Women's Underperformance in Math". In: *Psychological Science* 19.2 (2008), pp. 168–175. DOI: 10.1111/j.1467-9280.2008.02063.x.

[224] J. C. Croizet, G. Despres, M. E. Gauzins, P. Huguet, J. P. Leyens, and A. Meot. "Stereotype threat undermines intellectual performance by triggering a disruptive mental load". In: *Personality and Social Psychology Bulletin* 30.6 (2004).

[225] J. Keller and D. Dauenheimer. "Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math performance". In: *Personality and Social Psychology Bulletin* 29.3 (2003). DOI: 10.1177/0146167202250218.

[226] Charles Stangor, Christine Carr, and Lisa Kiang. "Activating Stereotypes Undermines Task Performance Expectations". In: *Journal of Personality and Social Psychology* 75.5 (1998), pp. 1191–1197. DOI: 10.1037/0022-3514.75.5.1191.

[227] J. Thomas Kellow and Brett D. Jones. "The Effects of Stereotypes on the Achievement Gap: Reexamining the Academic Performance of African American High School Students". In: *Journal of Black Psychology* 34.1 (2008), pp. 94–120.

[228] Calvin Graham, Robert W. Baker, and Seymour Wapner. "Prior interracial experience and Black student transition into predominantly White colleges". In: *Journal of Personality and Social Psychology* 47.5 (1984), pp. 1146–1154. DOI: 10.1037/0022-3514.47.5.1146.

[229] Joshua Aronson, Carrie B. Fried, and Catherine Good. "Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence". In: *Journal of experimental social psychology* 38.2 (2002). DOI: 10.1006/jesp.2001.1491.

[230] Carol S. Dweck, Chi-yue Chiu, and Ying-yi Hong. "Implicit Theories and Their Role in Judgments and Reactions: A Word From Two Perspectives". In: *Psychological Inquiry* 6.4 (1995), pp. 267–285. DOI: 10.1207/s15327965pli0604$\backslash${\textunderscore}1.

[231] Claude M. Steele. "The Psychology of Self-Affirmation: Sustaining the Integrity of the Self". In: ed. by Leonard Berkowitz. Vol. 21. Advances in Experimental Social Psychology. Academic Press, 1988, pp. 261–302. DOI: 10.1016/S0065-2601(08)60229-4.

[232] Claude M. Steele. "Name-Calling and Compliance". In: *Journal of Personality and Social Psychology* 31.2 (1975), pp. 361–369. DOI: 10.1037/h0076291.

[233] Amy MCQueen and WilliamM. P. Klein. "Experimental manipulations of self-affirmation: A systematic review". In: *Self & Identity* 5.4 (2006), pp. 289–354. DOI: 10.1080/15298860600805325.

[234] David K. Sherman and Geoffrey L. Cohen. "The psychology of self-defense: Self-affirmation theory". In: *Advances in Experimental Social Psychology* 38 (2006). DOI: 10.1016/S0065-2601(06)38004-5.

[235] Geoffrey L. Cohen, Julio Garcia, Nancy Apfel, and Allison Master. "Reducing the Racial Achievement Gap: A Social-Psychological Intervention". In: *Science* 313.5791 (2006), p. 1307.

[236] Geoffrey L. Cohen, Julio Garcia, Valerie Purdie-Vaughns, Nancy Apfel, and Patricia Brzustoski. "Recursive Processes in Self-Affirmation: Intervening to Close the Minority Achievement Gap". In: *Science* 324.5925 (2009), p. 400.

[237] David K. Sherman, Geoffrey L. Cohen, Leif D. Nelson, A. David Nussbaum, Debra P. Bunyan, and Julio Garcia. *Affirmed Yet Unaware: Exploring the Role of Awareness in the Process of Self-Affirmation*. 2009. DOI: 10.1037/a0015451.

[238] Natasha K. Bowen, Kate M. Wegmann, and Kristina C. Webber. "Enhancing a Brief Writing Intervention to Combat Stereotype Threat among Middle-School Students". In: *Journal of Educational Psychology* 105.2 (2013), pp. 427–435.

[239] Katherine Woolf, I. Chris McManus, Deborah Gill, and Jane Dacre. "The effect of a brief social intervention on the examination results of UK medical students: a cluster randomised controlled trial". In: *BMC Medical Education* 9 (2009). DOI: 10.1186/1472-6920-9-35.

[240] Judith M. Harackiewicz, Elizabeth A. Canning, Yoi Tibbetts, Cynthia J. Giffen, Seth S. Blair, Douglas I. Rouse, and Janet S. Hyde. "Closing the Social Class Achievement Gap for First-Generation Students in Undergraduate Biology". In: *Journal of Educational Psychology* 106.2 (2014), pp. 375–389.

[241] E.D.B. Riggle, K. A. Gonzalez, S. S. Rostosky, and W. W. Black. "Cultivating Positive LGBTQA Identities: An Intervention Study with College Students: Cultivating Positive LGBTQA Identities: An Intervention Study with College Students". In: *Journal of LGBT Issues in Counseling* 8.3 (2014), pp. 264–281. DOI: 10.1080/15538605.2014.933468.

[242] Gregory M. Walton, Christine Logel, Jennifer M. Peach, Steven J. Spencer, and Mark P. Zanna. "Two Brief Interventions to Mitigate a Chilly Climate Transform Women's Experience, Relationships, and Achievement in Engineering". In: *Journal of Educational Psychology* 107.2 (2015), pp. 468–485.

[243] Dominique N. Bradley, Evan P. Crawford, and Sara E. Dahill-Brown. "Defining and assessing FoI in a large-scale randomized trial: Core components of values affirmation". In: *Studies in Educational Evaluation* 49 (2016), pp. 51–65. DOI: 10.1016/j.stueduc.2016.04.002.

[244] Geoffrey Borman. "Examination of a Self-Affirmation Intervention in St. Paul Public Schools". In: *The Senior Urban Education Research Fellowship Series* IX (2012).

[245] Thomas S. Dee. "Social Identity and Achievement Gaps: Evidence from an Affirmation Intervention". In: *Journal of Research on Educational Effectiveness* 8.2 (2015), pp. 149–168.

[246] Geoffrey D. Borman, Jeffrey Grigg, and Paul Hanselman. "An Effort to Close Achievement Gaps at Scale Through Self-Affirmation". In: *Educational Evaluation and Policy Analysis* 38.1 (2016), pp. 21–42. DOI: 10.3102/0162373715581709.

[247] Hannah Jordt, Sarah L. Eddy, Riley Brazil, Ignatius Lau, Chelsea Mann, Sara E. Brownell, Katherine King, and Scott Freeman. "Values Affirmation Intervention Reduces Achievement Gap between Underrepresented Minority and White Students in Introductory Biology Classes". In: *CBE - Life Sciences Education* 16.3 (2017).

[248] Shanda Lauer, Jennifer Momsen, Erika Offerdahl, Mila Kryjevskaia, Warren Christensen, and Lisa Montplaisir. "Stereotyped: Investigating Gender in Introductory Science Courses". In: *CBE - Life Sciences Education* 12.1 (2013), pp. 30–38.

[249] Annie Franco, Neil Malhotra, and Gabor Simonovits. "Publication bias in the social sciences: Unlocking the file drawer". In: *Science* 345.6203 (2014), pp. 1502–1505. DOI: 10.1126/science.1255484.

[250] Jeffrey Mervis. "Research Transparency. Why null results rarely see the light of day". In: *Science* 345.6200 (2014), p. 992. DOI: 10.1126/science.345.6200.992.

[251] D. Fanelli. "Negative results are disappearing from most disciplines and countries: Negative results are disappearing from most disciplines and countries". In: *Scientometrics* 90.3 (2012), pp. 891–904. DOI: 10.1007/s11192-011-0494-7.

[252] D. Chavalarias, J. D. Wallach, A.H.T. Li, and J.P.A. Ioannidis. "Evolution of reporting P values in the biomedical literature, 1990-2015: Evolution of reporting P values in the biomedical literature, 1990-2015". In: *JAMA - Journal of the American Medical Association* 315.11 (2016), pp. 1141–1148. DOI: 10.1001/jama.2016.1952.

[253] R. Rosenthal. "The file drawer problem and tolerance for null results: The file drawer problem and tolerance for null results". In: *Psychological Bulletin* 86.3 (1979), pp. 638–641. DOI: 10.1037/0033-2909.86.3.638.

[254] Luke D. Conlin, Eric Kuo, and Nicole R. Hallinen. "Nothing's plenty: The significance of null results in physics education research". In: *arXiv e-prints* (2018), arXiv:1810.10071.

[255] T. D. Sterling, W. L. Rosenbaum, and J. J. Weinkam. "Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa". In: *The American Statistician* 49.1 (1995), p. 108. DOI: 10.2307/2684823.

[256] *Replicating a self-affirmation intervention to address gender differences: Successes and challenges.* Vol. 1413. 1. 2012, pp. 231–234. DOI: 10.1063/1.3680037.

[257] Moses Rifkin. "Addressing Underrepresentation: Physics Teaching for All". In: *Physics Teacher* 54.2 (2016), pp. 72–75.

[258] Zahra Hazari, Geoff Potvin, Robynne M. Lock, Florin Lung, Gerhard Sonnert, and Philip M. Sadler. "Factors that affect the physical science career interest of female students: Testing five common hypotheses". In: *Phys. Rev. ST Phys. Educ. Res.* 9.2 (2013), p. 020115. DOI: 10.1103/PhysRevSTPER.9.020115.

[259] E. A. Eschenbach, M. Virnoche, E. M. Cashman, S. M. Lord, and M. M. Camacho. "Proven practices that can reduce stereotype threat in engineering education: A literature review". In: *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*. 2014, pp. 1–9. DOI: 10.1109/FIE.2014.7044011.

[260] Lauren Aguilar, Greg Walton, and Carl Wieman. "Psychological insights for improved physics teaching". In: *Physics Today* 67.5 (2014). DOI: 10.1063/PT.3.2383.

[261] Jennifer L. Docktor and José P. Mestre. "Synthesis of discipline-based education research in physics". In: *Phys. Rev. ST Phys. Educ. Res.* 10.2 (2014), p. 020119. DOI: 10.1103/PhysRevSTPER.10.020119.

[262] Barry W Fitzgerald. "Using superheroes such as Hawkeye, Wonder Woman and the Invisible Woman in the physics classroom". In: *Physics Education* 53.3 (2018), p. 035032. DOI: 10.1088/1361-6552/aab442.

[263] Nina Abramzon, Patrice Benson, Edmund Bertschinger, Susan Blessing, Geraldine L. Cochran, Anne Cox, Beth Cunningham, Jessica Galbraith-Frew, Jolene Johnson, Leslie Kerby, Elaine Lalanne, Christine O'Donnell, Sara Petty, Sujatha Sampath, Susan Seestrom, Chandralekha Singh, Cherrill Spencer, Kathryne Sparks Woodle, and Sherry Yennello. "Women in physics in the United States: Recruitment and retention". In: *AIP Conference Proceedings* 1697.1 (2015), p. 060045. DOI: 10.1063/1.4937692.

[264] T. Scott, A. Gray, and P. Yates. "A controlled comparison of teaching methods in first-year university physics". In: *Journal of the Royal Society of New Zealand* 43.2 (2013), pp. 88–99. DOI: 10.1080/03036758.2012.658816.

[265] Jeffrey M. Valla and Wendy M. Williams. "Increasing Achievement and Higher-Education Representation of Under-Represented Groups in Science, Technology, Engineering, and Mathematics Fields: A Review of Current K-12 Intervention Programs". In: *Journal of women and minorities in science and engineering* 18.1 (2012), pp. 21–53. DOI: 10.1615/JWomenMinorScienEng.2012002908.

[266] Rachel Henderson, Gay Stewart, John Stewart, Lynnette Michaluk, and Adrienne Traxler. "Exploring the Gender Gap in the Conceptual Survey of Electricity and Magnetism". In: *Physical Review Physics Education Research* 13.2 (2017), pp. 020114–1.

[267] Kimberley Kreutzer and Andrew Boudreaux. "Preliminary investigation of instructor effects on gender gap in introductory physics". In: *Phys. Rev. ST Phys. Educ. Res.* 8.1 (2012), p. 010120. DOI: 10.1103/PhysRevSTPER.8.010120.

[268] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman. "New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey". In: *Phys. Rev. ST Phys. Educ. Res.* 2.1 (2006), p. 010101. DOI: 10.1103/PhysRevSTPER.2.010101.

Appendices

Appendix A

Physics 100 Diagnostic Test

The purpose of this diagnostic instrument is to identify students who would benefit from extra instruction in how to approach physics problems. This extra instruction takes the form of the 1-2 hour class PHYS 100. This diagnostic also helps to assess some previous physics and mathematics knowledge.

Select the answer you think best to each of these multiple-choice questions. You will probably need a pencil and scratch paper to work out some of the problems. You may also use a calculator. If you have no idea how to solve the problem and would just be making a random guess, select "I can't eliminate any choices." This gives us a more accurate assessment of your physics/math preparation than having you randomly guess answers. If you can eliminate at least one or two choices, select the answer that you think is best.

INTRODUCTION These first three questions tell us a little about your background.

I. How many years of high school physics did you take?

a. 0 years
b. 1 year
c. 2 or more years

II. If you took any physics in high school, how would you rate the quality of the class?

a. Good
b. Fair
c. Poor
d. I didn't take any physics in high school

III. Do you think you will have trouble passing PHYS 211?

a. Yes
b. No, but I will probably take PHYS 100 if it's recommended
c. No, and I probably won't probably take PHYS 100 if it's recommended

The remaining questions test your math/physics preparedness.

1. A ball is thrown straight up in the air with a velocity of 40 m/s.
Neglecting air resistance, how long will the ball be in the air?

a. Time = 3.57 s
b. Time = 8.16 s
c. Time = 12.46 s
d. Time = 20 s
e. Time = 40 s
f. I can't eliminate any choices

2. Car A leaves town heading east with a constant speed of 20 miles/hr. An hour later car B leaves town heading east as well, with a constant speed of 25 miles/hr.
How long after car B leaves will it intercept car A?

a. Half an hour later
b. Two hours later
c. Three hours later
d. Four hours later
e. Five hours later
f. I can't eliminate any choices

3. A child runs at speed v_c, while an adult runs at speed v_a. If the adult gives the child a head start of distance d, how long will it take for the adult to catch the child?
Express your answer in terms of v_a, v_c, and d.

a. $d/(v_a + v_c)$
b. $d/(v_a - v_c)$
c. $(v_c + d)/v_a$
d. d/v_a
e. I can't eliminate any choices

4. A ball is orbiting counterclockwise at 100 revolutions per minute (rpm) around a circle of radius 10 cm. The center of the circle is at the x-y origin (0, 0).
At t = 0, the ball is at (10 cm, 0).
When does the ball first reach the y axis?

a. t = 0.15 sec
b. t = 0.20 sec
c. t = 0.25 sec
d. t = 0.30 sec
e. I can't eliminate any choices

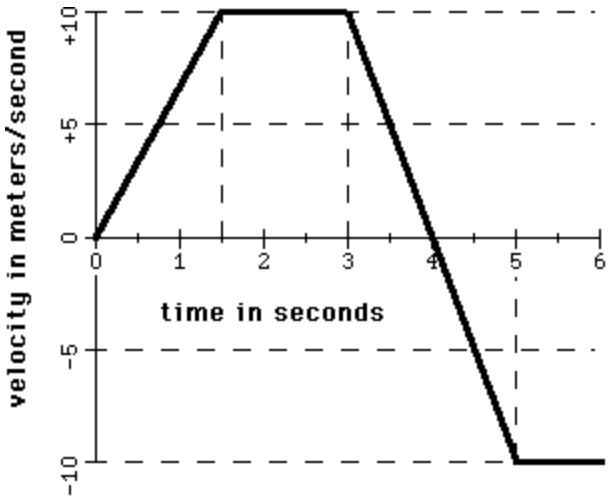


Figure A.1: Image for Question 5.

5. This graph shows the velocity of an object versus time.
At which time does the object change its direction of motion?

a. 1.5 s
b. 3 s
c. 4 s
d. 5 s
e. None of the above

6. A car drives east for two hours at 40 miles per hour. After stopping for two hours, it then drives east for another three hours at 60 miles per hour. Which of the following graphs correctly represents the distance versus time for this motion?

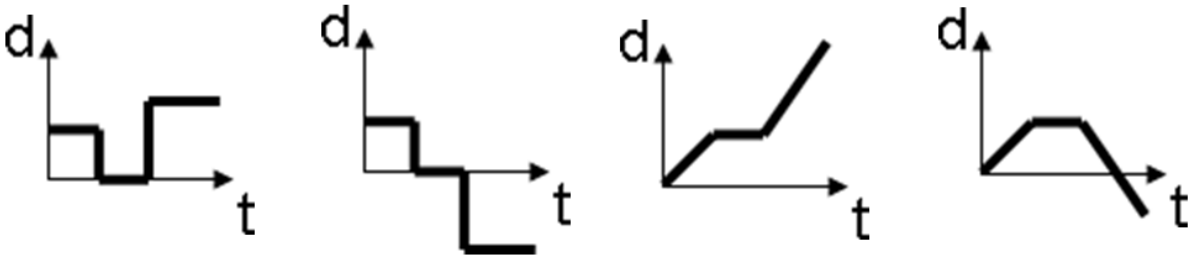


Figure A.2: Question 6, choices a-d.

e. I can't eliminate any choices

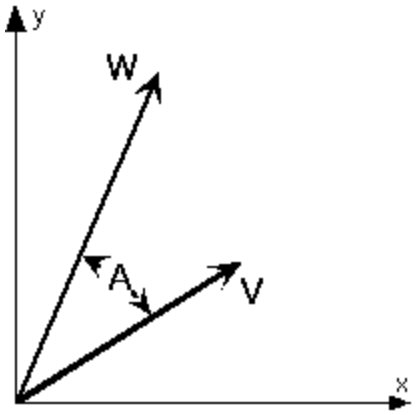


Figure A.3: Image for Question 7.

The next two questions deal with the following situation:

Here we have two vectors **V** and **W**. The angle between these vectors is A.

7. What is the component of **V** parallel to **W** in terms of A and the magnitudes of **V** and **W**?

a. V
b. W
c. $V\sin(A)$
d. $V\cos(A)$
e. $W\sin(A)$
f. $W\cos(A)$
g. I can't eliminate any choices

8. What is the component of **V** perpendicular to **W** in terms of A and the magnitudes of **V** and **W**?

a. V
b. W
c. $V\sin(A)$
d. $V\cos(A)$
e. $W\sin(A)$
f. $W\cos(A)$
g. I can't eliminate any choices

The next two questions deal with the following situation:

The two vectors **A** and **B** are defined as: **A** = (3 m, 4 m) and **B** = (5 m, -3 m), where the first number is the x-component of the vector, and the second number is the y-component of the vector.

9. What is the magnitude of vector **A**?

a. 7 m
b. 4 m
c. (8 m, 1 m)
d. 5 m
e. (3 m, 4 m)
f. I can't eliminate any choices

10. What is **A** + **B**?

a. (8 m, 1 m)
b. 7 m
c. (3 m, 1 m)
d. 10.83 m
e. 5.83 m
f. (7 m, 2 m)
g. I can't eliminate any choices

11. In the year 2003 the Acme Company has twice as many employees as managers. During the next 5 years, they hire twelve employees and fire two managers.
Which of the following equations is correct for the current number of managers (NM_{2008}) and the current number of employees (NE_{2008})?

a. $NM_{2008} + 2 = 2(NE_{2008} - 12)$
b. $NM_{2008} - 2 = 2(NE_{2008} + 12)$
c. $2(NM_{2008} + 2) = NE_{2008} - 12$
d. $2(NM_{2008} - 2) = NE_{2008} + 12)$
e. I can't eliminate any choices

PROBLEM VIII

The next two questions both use the following information:
The force on an object in uniform circular motion is given by $F = mv^2/r$.

12. Consider two equal-mass objects moving around the same circle, one with speed v_1 and one with speed v_2, where $v_1 = 2\ v_2$. What is the ratio of the forces on these two objects?

a. $F_1/F_2 = 1$
b. $F_1/F_2 = 2$
c. $F_1/F_2 = 4$
d. $F_1/F_2 = 1/2$
e. $F_1/F_2 = 1/4$

13. Now the two objects move at the same speed around the same circle but have different masses m_1 and m_2, where $m_1 = 2\ m_2$. What is the ratio of the forces on the two objects?

a. $F_1/F_2 = 1$
b. $F_1/F_2 = 2$
c. $F_1/F_2 = 4$
d. $F_1/F_2 = 1/2$
e. $F_1/F_2 = 1/4$

Appendix B

Clinical Trial Materials

The following are the materials seen by students who participated in the clinical trials described in Chapters 3 and 4.

Mastery Exercises

One version of each level is provided in this appendix, though students had access to four when working through the mastery exercises. Variations among versions of these questions consisted of changing the signs of the plates, shells, and slabs, and asking about different regions. Numerical values were randomized within each version.

The breakup of the questions shown here is from the second implementation; there were 2 levels in the first implementation and 4 levels in the second, but the questions are the same. We saw in our first implementation that having too many questions in a single level was frustrating to students so Level 1 was split into two smaller levels (Levels 1 and 2 shown here) and Level 2 became smaller Levels 3 and 4 (shown here).

Level 1: Superposition of Electric Fields

An infinite sheet of charge is located in the y-z plane at x=0 and has uniform negative charge density $-\sigma_1\ C/m^2$. Another infinite sheet of charge with uniform positive charge density $+\sigma_2\ C/m^2$ is located at x=2m, as shown.

Note that the symbols σ_1 and σ_2 represent the magnitude of the charge density, and are always positive numbers when used below. In the answers, ϵ_o is equal to $8.85 \times 10^{-12}\ C^2/Nm^2$.

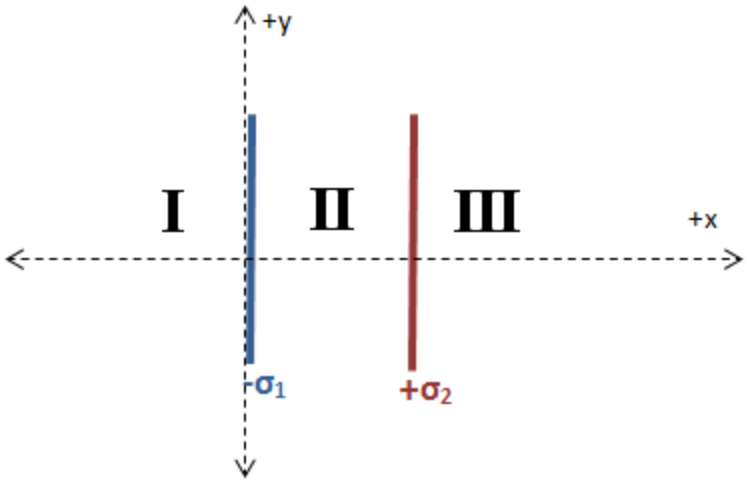


Figure B.1: Figure given with prompt for Level 1, Question 1.

1. What is $E_x\,(II)$, the x-component of the electric field in region II?

a. $\dfrac{(\sigma_1+\sigma_2)}{2\epsilon_o}$ N/C
b. $\dfrac{(\sigma_1-\sigma_2)}{2\epsilon_o}$ N/C
c. $\dfrac{(\sigma_2-\sigma_1)}{2\epsilon_o}$ N/C
d. $-\dfrac{(\sigma_1+\sigma_2)}{2\epsilon_o}$ N/C
e. $\dfrac{\sigma_1}{2\epsilon_o}$ N/C
f. $-\dfrac{\sigma_1}{2\epsilon_o}$ N/C
g. 0 N/C

The next three questions concern the situation below.

An infinite sheet of charge is located in the y-z plane at x=0 and has uniform negative charge density $-\sigma_1\ C/m^2$. Another infinite sheet of charge with uniform positive charge density $+\sigma_2\ C/m^2$ is located at x=5m. In between these two sheets is a conducting slab with uniform positive charge density $+\sigma_3\ C/m^2$, as shown.

Note that the symbols σ_1, σ_2, and σ_3 represent the magnitude of the charge density, and are always positive numbers when used below. In the answers, ϵ_o is equal to $8.85\times10^{-12}\ C^2/Nm^2$.

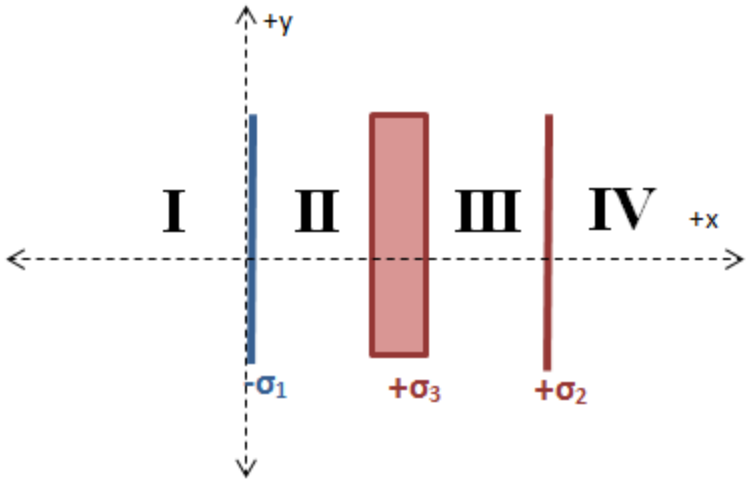


Figure B.2: Figure given with prompt for Level 1, Questions 2-4.

2. What is $E_x\,(III)$, the x-component of the electric field in region III?

a. $\frac{(\sigma_1+\sigma_2+\sigma_3)}{2\epsilon_o}$ N/C
b. $-\frac{(\sigma_1+\sigma_2+\sigma_3)}{2\epsilon_o}$ N/C
c. $\frac{(\sigma_1-\sigma_2-\sigma_3)}{2\epsilon_o}$ N/C
d. $-\frac{(\sigma_1-\sigma_2-\sigma_3)}{2\epsilon_o}$ N/C
e. $\frac{\sigma_1}{2\epsilon_o}$ N/C
f. $-\frac{\sigma_1}{2\epsilon_o}$ N/C
g. 0 N/C

3.What is $E_x\,(insideslab)$, the x-component of the electric field inside the conducting slab?

a. $\frac{(\sigma_1+\sigma_2)}{2\epsilon_o}$ N/C
b. $\frac{(\sigma_1-\sigma_2)}{2\epsilon_o}$ N/C
c. $\frac{(\sigma_2-\sigma_1)}{2\epsilon_o}$ N/C
d. $-\frac{(\sigma_1+\sigma_2)}{2\epsilon_o}$ N/C
e. $\frac{\sigma_2}{2\epsilon_o}$ N/C
f. $-\frac{\sigma_2}{2\epsilon_o}$ N/C
g. 0 N/C

4. Using the same setup as above, suppose that $\sigma_1 = 10\ C/m^2$, $\sigma_2 = 5\ C/m^2$, and $\sigma_3 = 8\ C/m^2$. What would be the value of $E_x\,(III)$,the x-component of the electric field in region III?

a. 7.34×10^{11} N/C
b. -7.34×10^{11} N/C
c. 1.3×10^{12} N/C
d. -1.3×10^{12} N/C
e. 3.95×10^{11} N/C
f. -3.95×10^{11} N/C
g. 0 N/C

Level 2: Planar Surface Charges

The next two questions concern the situation shown below.

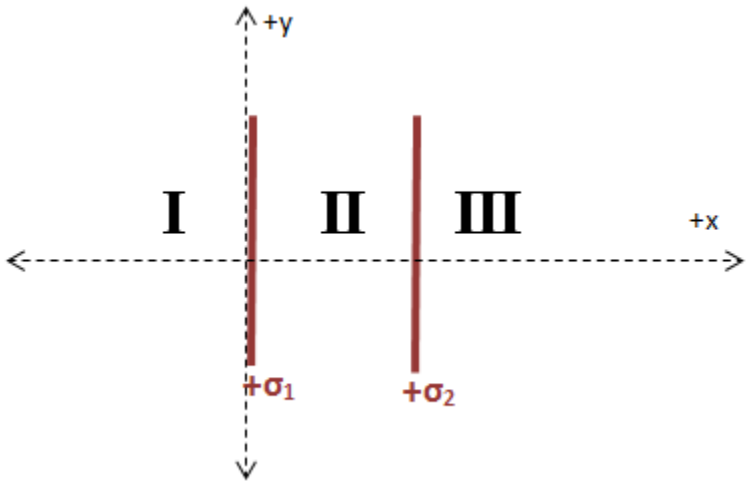


Figure B.3: Figure given with prompt for Level 2, Questions 1-2.

An infinite sheet of charge is located in the y-z plane at x=0 and has uniform positive charge density $+\sigma_1$ C/m^2. Another infinite sheet of charge with uniform positive charge density $+\sigma_2$ C/m^2 is located at x=2m.

Note that the symbols σ_1 and σ_2 represent the magnitude of the charge density, and are always positive numbers when used below. In the answers, ϵ_o is equal to 8.85×10^{-12} C^2/Nm^2.

1. What relationship between σ_1 and σ_2 could result in a negative x-component of the electric field in region II?

Remember that σ_1 and σ_2 are the magnitudes of the charge.

a. $\sigma_1 = \sigma_2$
b. $\sigma_1 > \sigma_2$
c. $\sigma_1 < \sigma_2$
d. There are no values of σ_1 and σ_2 that could result in a negative x-component of E in region II.

2. Suppose $\sigma_2 = 4\ \mu C/m^2$ and the x-component of the electric field in region II, $E_x\,(II)$, is equal to 8.5×10^4 N/C. Then, what is σ_1?

a. 1.5 $\mu C/m^2$
b. 4.8 $\mu C/m^2$
c. 3.2 $\mu C/m^2$
d. 2.5 $\mu C/m^2$
e. 5.5 $\mu C/m^2$
f. 0 $\mu C/m^2$

The next question concerns the NEW situation shown below.

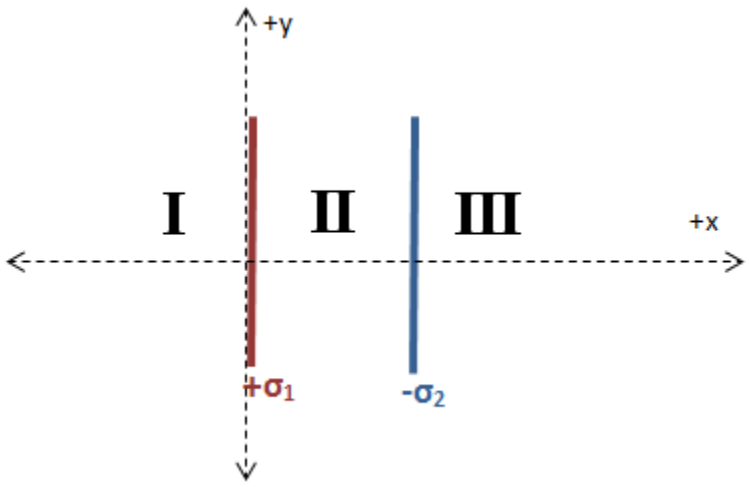


Figure B.4: Figure given with prompt for Level 2, Question 3.

An infinite sheet of charge is located in the y-z plane at x=0 and has uniform positive charge density $+\sigma_1\ C/m^2$. Another infinite sheet of charge with uniform negative charge density $-\sigma_2\ C/m^2$ is located at x=2m.

Note that the symbols σ_1 and σ_2 represent the magnitude of the charge density, and are always positive numbers when used below. In the answers, ϵ_o is equal to $8.85 \times 10^{-12}\ C^2/Nm^2$.

3. What relationship between σ_1 and σ_2 could result in the electric field in region I being equal to 0?

a. $\sigma_1 = \sigma_2$
b. $\sigma_1 > \sigma_2$
c. $\sigma_1 < \sigma_2$
d.There are no values of σ_1 and σ_2 that could result in E=0 in region I.

The next question concerns the NEW situation shown below.

An infinite sheet of charge is located in the y-z plane at x=0 and has uniform positive charge density $\sigma_1\ C/m^2$. Another infinite sheet of charge with uniform positive charge density $+\sigma_2\ C/m^2$ is located at

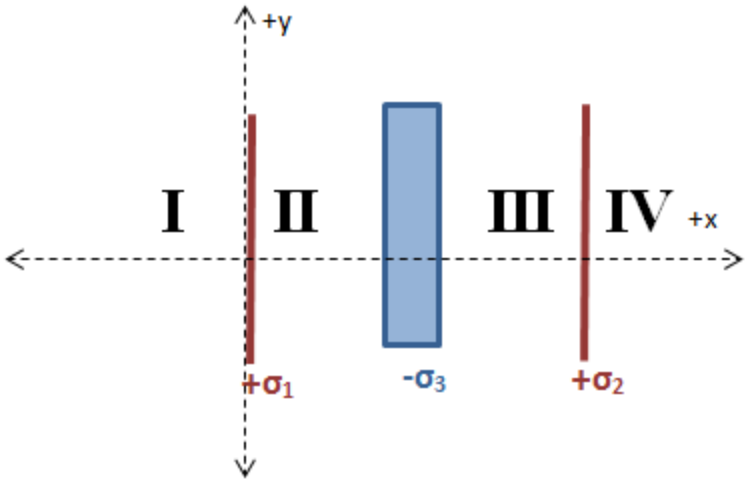


Figure B.5: Figure given with prompt for Level 2, Question 4.

x=5m. In between these two sheets is a conducting slab with uniform negative charge density $-\sigma_3$ C/m^2, as shown.

Note that the symbols σ_1, σ_2, and σ_3 represent the magnitude of the charge density, and are always positive numbers when used below. In the answers, ϵ_o is equal to 8.85×10^{-12} C^2/Nm^2.

4. What value of σ_1 would produce an electric field equal to 0 in region I?

a. $(\sigma_2 + \sigma_3)$
b. $-(\sigma_2 + \sigma_3)$
c. $(\sigma_2 - \sigma_3)$
d. $(\sigma_3 - \sigma_2)$
e. 0

Level 3: Fields and Potentials from Spherical Charges (Single Integral)

An electric field is defined through all space by E(x) = 3x $\hat{i}$ N/C where x is measured in meters.

1. What is V(x=8m) − V(x=12m)?

a. 12 V
b. −12 V
c. 120 V
d. −120 V
e. 0.125 V
f. −0.125 V

The next four questions refer to the situation below.

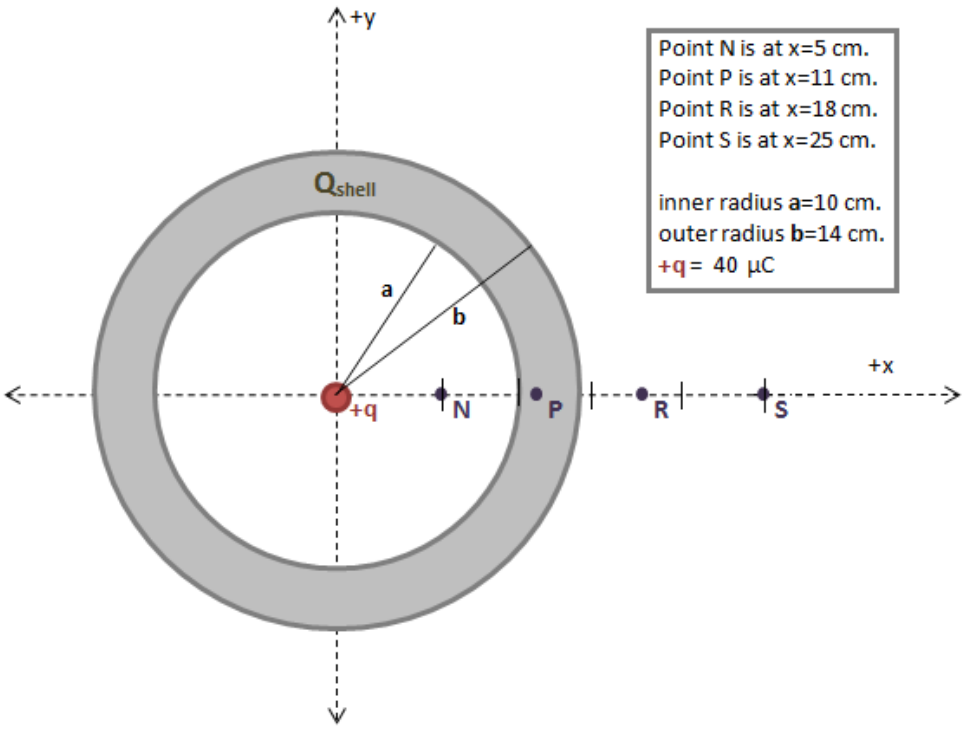


Figure B.6: Figure given with prompt for Level 3, Questions 2-5.

A positive charge, $q = 40\ \mu C$ is placed at the center of a charged spherical shell conductor, with charge Q_{shell}, as shown. The values or coordinates of relevant points are given in the diagram.

2. Given that the magnitude of the E-field at point **S** is 8.64×10^6 N/C, what is Q_{shell}?

a. $40\ \mu C$
b. $-40\ \mu C$
c. $20\ \mu C$
d. $-20\ \mu C$
e. $60\ \mu C$
f. $-60\ \mu C$
g. $80\ \mu C$
h. $-80\ \mu C$

For the next three questions, assume the shell has been charged to the given value.

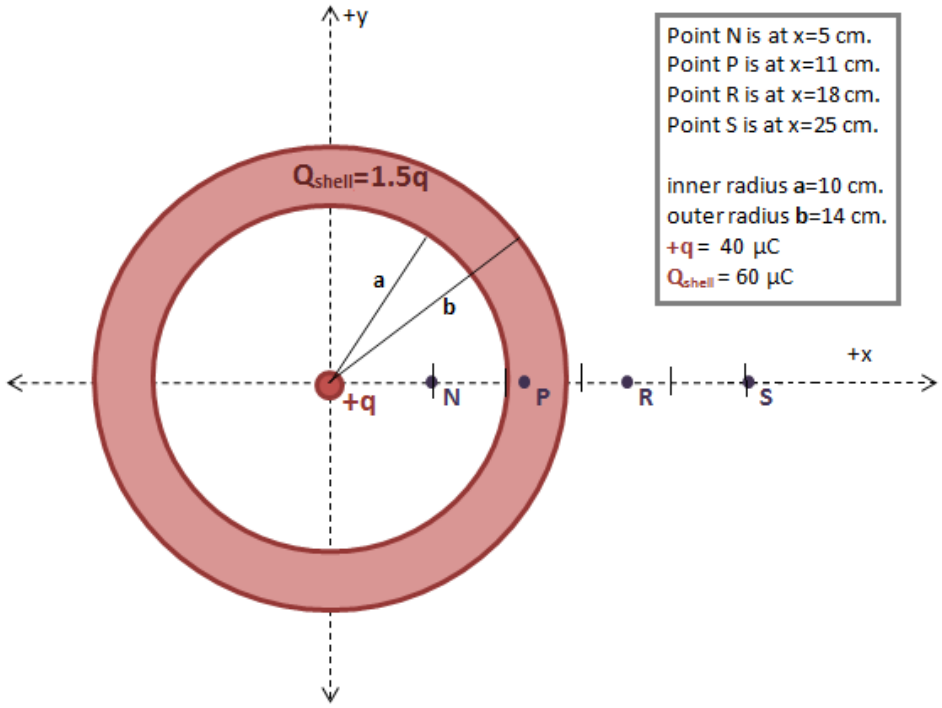


Figure B.7: Figure given with prompt for Level 3, Questions 3-5.

Now, suppose the shell has been charged to have $Q_{shell} = 1.5q = 60\ \mu C$.

3. What is E_x at point **R**, where x = 18 cm?

a. 1.67×10^7 N/C
b. -1.67×10^7 N/C
c. 2.78×10^7 N/C
d. -2.78×10^7 N/C
e. 1.11×10^7 N/C
f. -1.11×10^7 N/C
g. 0 N/C

3. What is E_x at point **P**, where x = 11 cm?

a. 2.98×10^7 N/C
b. -2.98×10^7 N/C
c. 2.25×10^7 N/C
d. -2.25×10^7 N/C
e. 3.6×10^7 N/C
f. -3.6×10^7 N/C
g. 0 N/C

4. What is V(R) − V(S)?

a. 1.4×10^6 V
b. -1.4×10^6 V
c. 3×10^6 V
d. -3×10^6 V
e. 8.4×10^6 V
f.-8.4×10^6 V
g. 0 V

Level 4: Fields and Potentials from Spherical Charges (Multiple Integrals)

For the next two questions, the electric field as a function of x is given by

$$E(x) = \begin{cases} -4x^2 & \hat{i}\ N/C, \quad 0 < x < 3m \\ 0 & N/C, \quad 3m < x < 6m \\ -24x & \hat{i}\ N/C \quad 6m < x < 12m \end{cases}$$

where x is measured in meters.

1. What is V(x=10m) − V(x=5m)?

a. 768 V
267 b. − 768 V
c. 120 V
d. − 120 V
e. 900 V
f. − 900 V

2. What is V(x=2m) −V (x=8m)?

a. 361.3 V
b. − 361.3 V
c. 176 V
d. − 176 V
e. 757.3 V
f. − 757.3 V

The next two questions refer to the situation below.

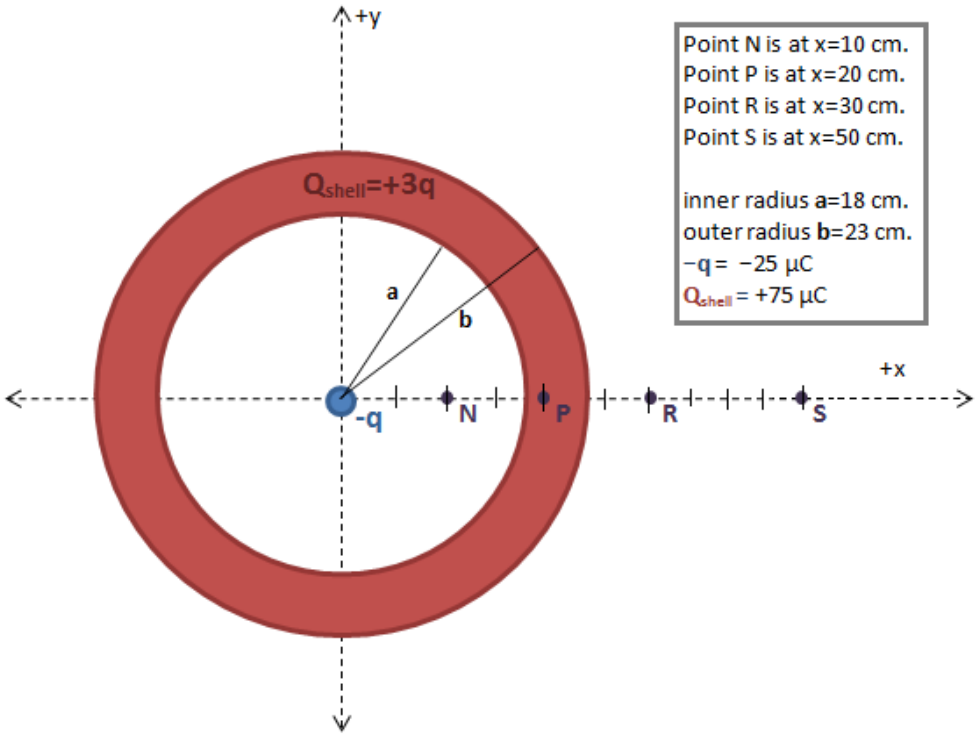


Figure B.8: Figure given with prompt for Level 4, Questions 3-4.

A negative charge, $-q = -25\ \mu C$ is placed at the center of a positively charged spherical shell conductor, with $Q_{shell} = +3q = +75\ \mu C$, as shown. The values or coordinates of relevant points are given in the diagram.

3. What is V(N) − V(S)?

a. 2.1×10^6 V
b. -2.1×10^6 V
c. 5.7×10^6 V
d. -5.7×10^6 V
e. 3.6×10^6 V
f. -3.6×10^6 V
g. 0 V

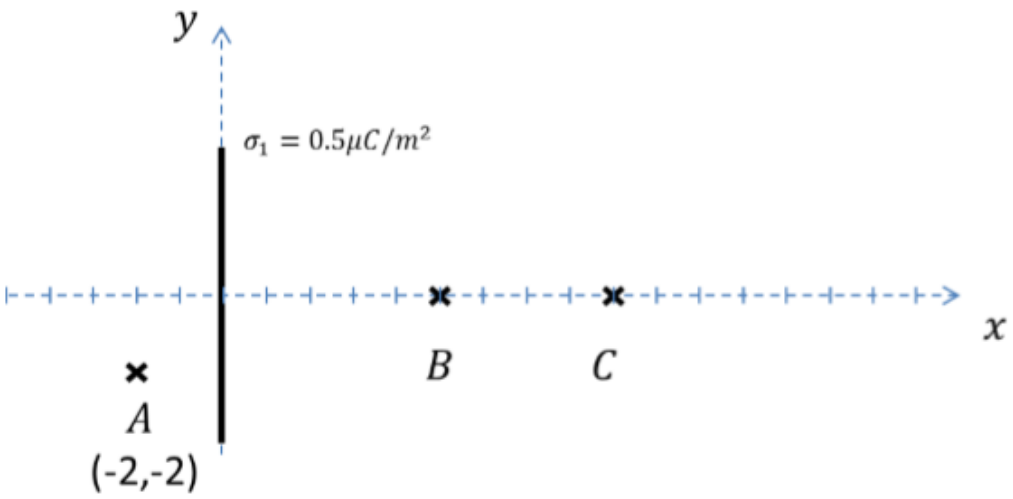
4. Suppose the charge at the origin is changed to $+q$. How does V(N) − V(S) change?

a. V(N) − V(S) increases.
b. V(N) − V(S) decreases.
c. V(N) − V(S) doesn't change.

Written Post-test

THE NEXT 2 QUESTIONS REFER TO THE SITUATION DESCRIBED BELOW

An infinite sheet of charge is located in the y-z plane at x = 0 and has uniform charge density $\sigma_1 = 0.5\ \mu C/m^2$. Point A is at (x,y)=(-2cm, -2cm). Point B is at (x,y)=(5 cm, 0 cm). Point C is at (x,y)=(9 cm, 0 cm).

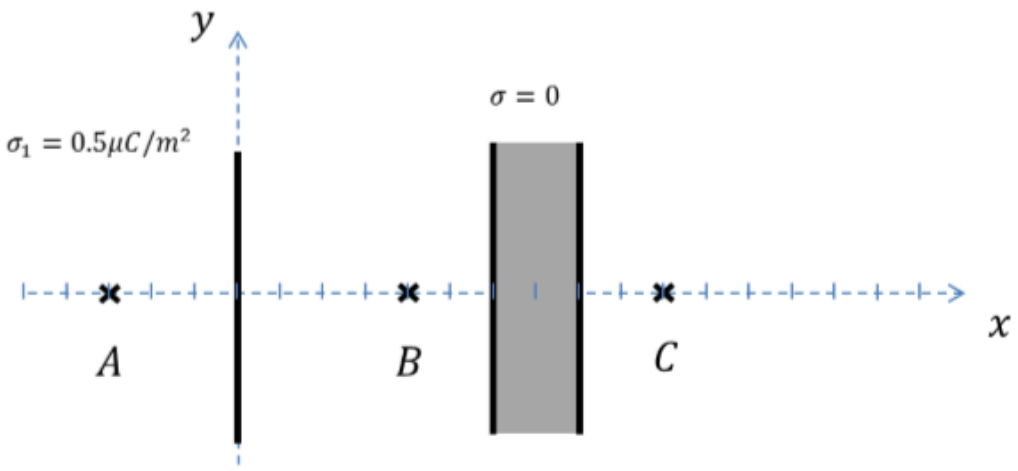


1) Calculate $V_B - V_C$

2) Calculate $V_A - V_B$

THE NEXT 2 QUESTIONS REFER TO THE SITUATION DESCRIBED BELOW

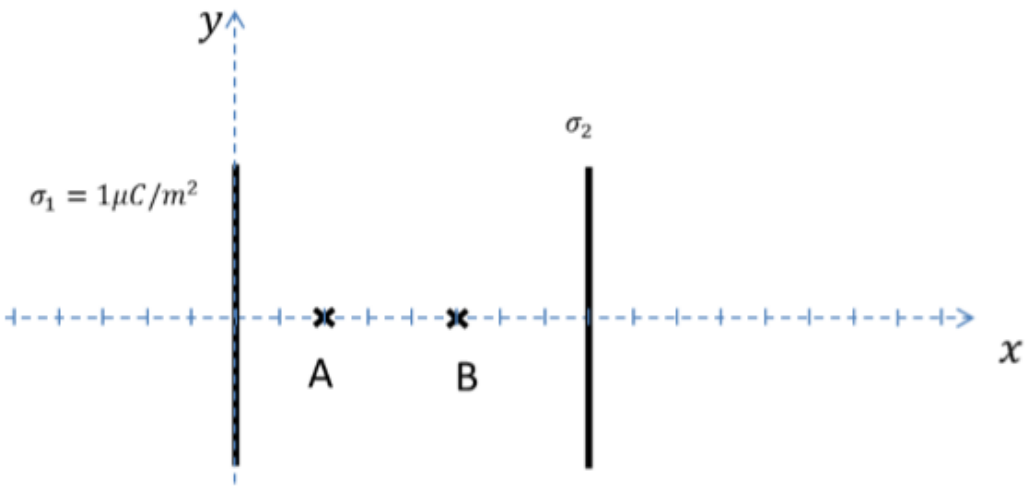
An infinite sheet of charge is located in the y-z plane at x = 0 and has uniform charge density $\sigma_1 = 0.5\ \mu C/m^2$. A 2cm wide conducting slab is placed with its left edge at x=6cm and its right edge at x=8cm. Point A is at (x,y)=(-3 cm,0 cm). Point B is at (x,y)=(4 cm, 0 cm). Point C is at (x,y)=(10 cm, 0 cm).



3) Calculate $V_B - V_C$

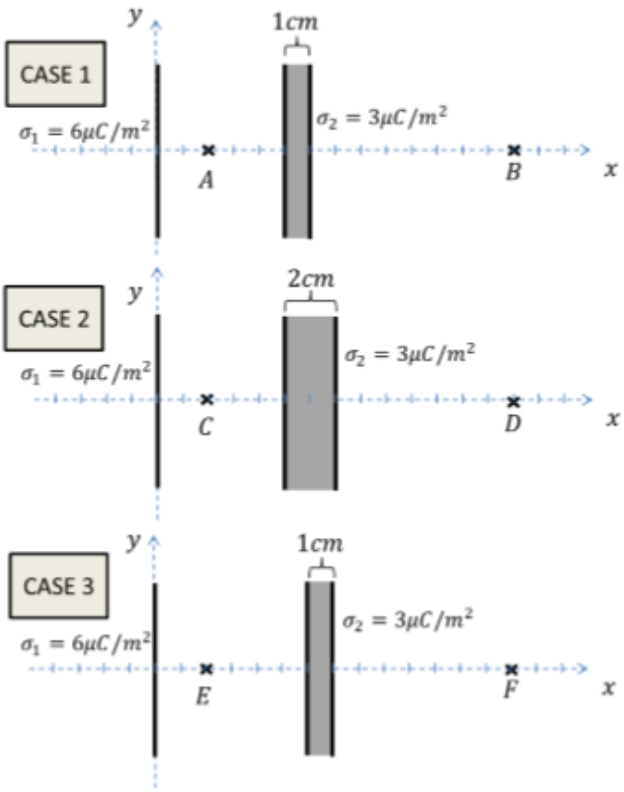
4) Calculate $V_A - V_C$

An infinite sheet of charge is located in the y-z plane at x = 0 and has uniform charge density $\sigma_1 = 1\ \mu C/m^2$. A second infinite sheet of charge is located in the y-z plane at x = 8 cm and has an unknown charge density σ_2. Point A is at (x,y)=(2 cm,0 cm). Point B is at (x,y)=(5 cm, 0 cm). The potential difference $V_A - V_b$ is known to be $15000\ V$.



5) What is σ_2?

Three situations are shown. In each case, an infinite sheet of charge is located in the y-z plane at x = 0 and has uniform positive charge density $\sigma_1 = 6\mu C/m^2$, and the conducting slab has uniform positive charge density $\sigma_2 = 3\mu C/m^2$. Points A, C, and E all lie at (x,y)=(2cm,0) on their respective coordinate axes. Points B,D, and F all lie at (x,y)=(14cm,0). In case 1, the charged conducting slab is 1cm wide and is placed with its left edge at x=5cm and its right edge at x=6cm. In case 2, the charged conducting slab is 2cm wide and is placed with its left edge at x=5cm and its right edge at x=7cm. In case 3, the charged conducting slab is 1cm wide and placed with its left edge at x=6cm and its right edge at x=7cm. The cases are diagrammed below.



6) Compare $|V_A - V_B|$ and $|V_C - V_D|$

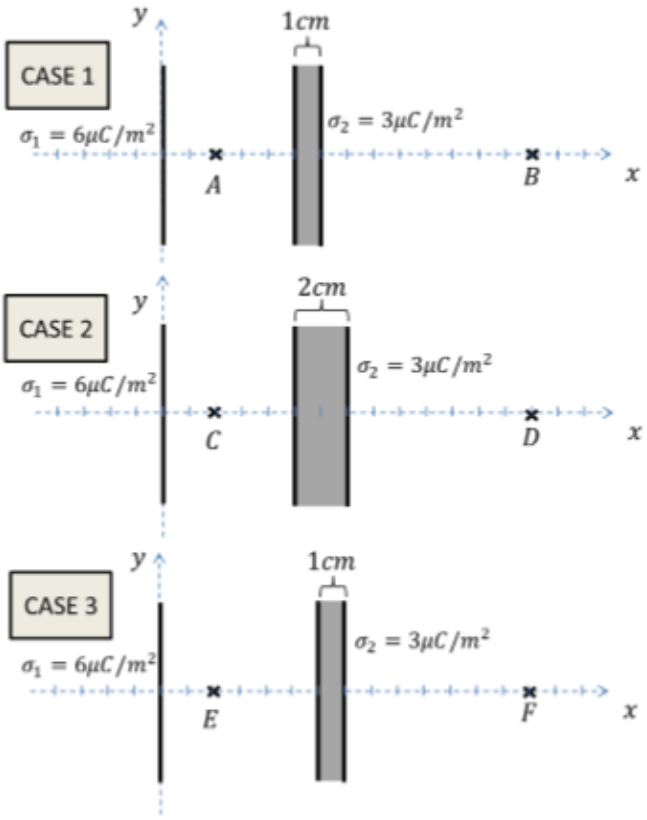
   a. $|V_A - V_B| > |V_C - V_D|$

   b. $|V_A - V_B| = |V_C - V_D|$

   c. $|V_A - V_B| < |V_C - V_D|$

Please explain your answer to the previous question

Three situations are shown. In each case, an infinite sheet of charge is located in the y-z plane at x = 0 and has uniform positive charge density $\sigma_1 = 6\mu C/m^2$, and the conducting slab has uniform positive charge density $\sigma_2 = 3\mu C/m^2$. Points A, C, and E all lie at (x,y)=(2cm,0) on their respective coordinate axes. Points B,D, and F all lie at (x,y)=(14cm,0). In case 1, the charged conducting slab is 1cm wide and is placed with its left edge at x=5cm and its right edge at x=6cm. In case 2, the charged conducting slab is 2cm wide and is placed with its left edge at x=5cm and its right edge at x=7cm. In case 3, the charged conducting slab is 1cm wide and placed with its left edge at x=6cm and its right edge at x=7cm. The cases are diagrammed below.



7) Compare $|V_A - V_B|$ and $|V_E - V_F|$

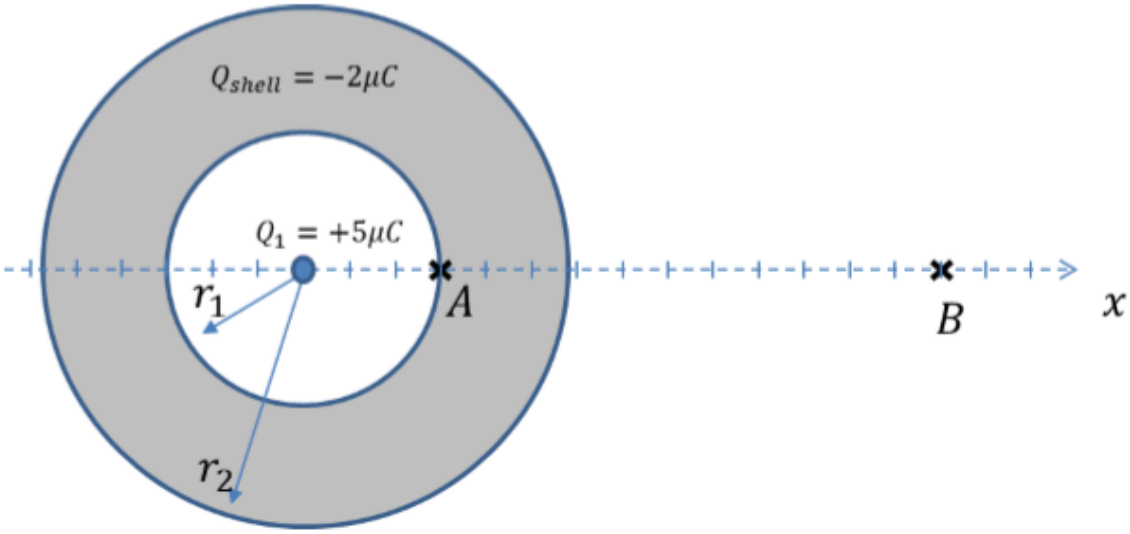
   a. $|V_A - V_B| > |V_E - V_F|$

   b. $|V_A - V_B| = |V_E - V_F|$

   c. $|V_A - V_B| < |V_E - V_F|$

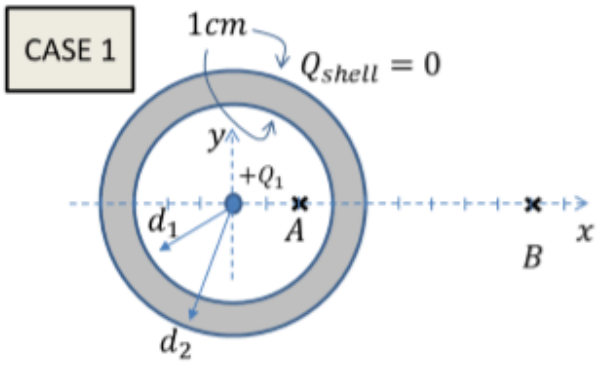
Please explain your answer to the previous question

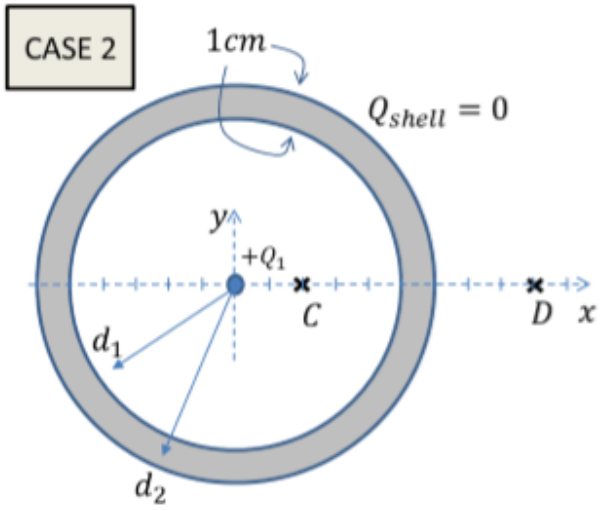
A point charge $Q_1 = +5\mu C$ is at the origin. A charged concentric conducting shell of inner radius r_1=3cm and outer radius r_2=6cm holds a charge $Q_{shell} = -2\mu C$. Point A is at (x,y)=(3cm,0cm). Point B is at (x,y)=(14cm,0).



8) Calculate $V_A - V_B$

A positive point charge Q_1 is placed at the origin. Points A and C all lie at (x,y)=(2cm,0) on their respective coordinate axes. Points B and D lie at (x,y)=(11cm,0). An uncharged 1cm thick concentric conducting spherical shell is placed outside the point charge. In case 1, the inner radius d_1=3cm, and d_2=4cm. In case 2, d_1=5cm, and d_2=6cm.





9) Compare V_{AB} and V_{CD}

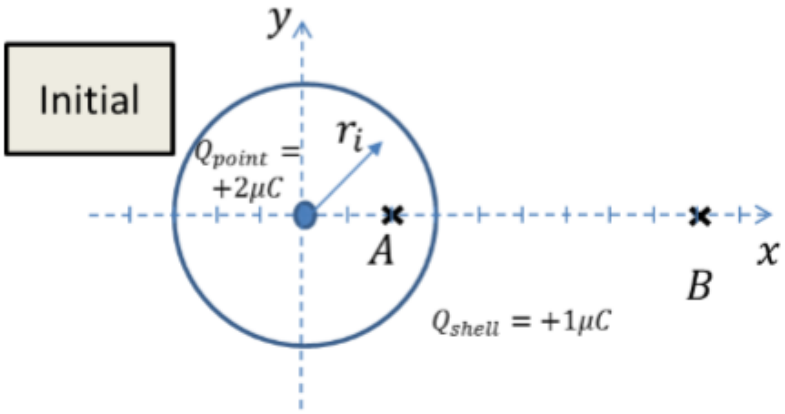
a. $|V_A - V_B| > |V_C - V_D|$

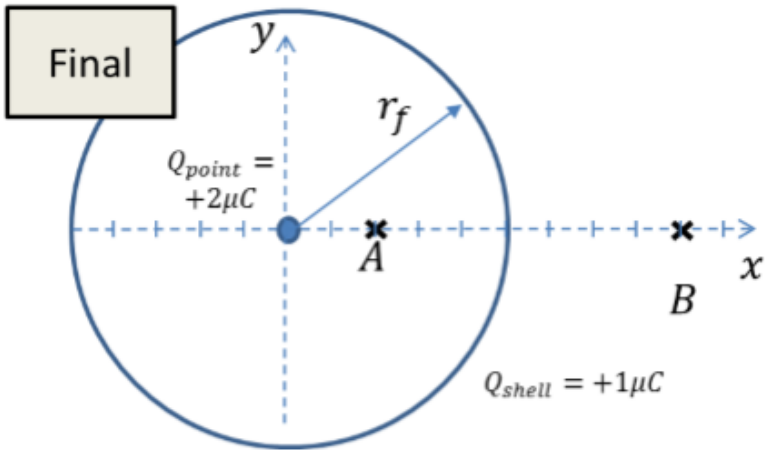
b. $|V_A - V_B| = |V_C - V_D|$

c. $|V_A - V_B| < |V_C - V_D|$

Please explain your answer to the previous question

A point charge $Q_{point} = +2\mu C$ is placed at the origin. The point charge is surrounded by a thin concentric conducting spherical shell with charge $Q_{shell} = +1\mu C$. Point A is at (x,y)=(2cm,0), and point B is at (x,y)=(11cm,0). Initially the radius of the shell is r_i. While the charge Q_{shell} is held constant, the radius is then increased to r_f, such that $r_f > r_i$.





10) When the radius increases, how does $|V_A - V_B|$ change?

a. $|V_A - V_B|$ increases

b. $|V_A - V_B|$ stays the same

c. $|V_A - V_B|$ decreases

Please explain your answer to the previous question

Appendix C

Physics 100 Homework Levels and Competencies (Fall 2014)

The following is a list of levels and corresponding competencies for the Physics 100 mastery homework from 2014.

Week 1: Kinematic definitions

Level 1: v-t graphs

1. To be able to translate a v vs t graph description of motion of a single object into a verbal description of motion.
2. To be able to identify a graphical description of motion (v vs t graph) of a single object that corresponds to a given verbal description of motion.

Level 2: x-t graphs

1. To be able to translate an x vs t graph description of motion of a single object into a verbal description of motion.
2. To be able to identify a graphical description of motion (x vs t graph) of a single object that corresponds to a given verbal description of motion.

Level 3: Multiple-body graphs

1. To be able to identify the x vs t graph that corresponds to a given verbal description of motion of two objects.
2. To be able to identify an algebraic equation of motion for one of the two objects.

3. To be able to determine some property relating to the combined motion of the objects.

Week 2: Constant acceleration and 1D relative motion

Level 1: Relative motion in 1D

1. To be able to calculate velocities of objects in a new reference frame, given these velocities in a common reference frame.
2. To be able to identify a plot of x vs t for two objects in the new reference frame.
3. To be able to calculate times or distances for specific events defined by the motion of the objects.

Level 2: 1D Constant acceleration 1

1. To be able to solve 1D constant acceleration (both free fall and non-free fall) problems involving one object that do not require the use of quadratic equations or coupled equations.

Level 3: 1D Constant acceleration 2

1. To be able to solve 1D constant acceleration (both free fall and non-free fall) problems involving one object that do not require the use of quadratic equations but can require the use of coupled equations.

Level 4: 1D Constant acceleration 3

1. To be able to solve numerical 1D constant acceleration (both free fall and non-free fall) problems involving two objects that can require the use of quadratic equations.

Level 5: 1D Constant acceleration 4

1. To be able to solve symbolic 1D constant acceleration (both free fall and non-free fall) problems involving two objects that can require the use of quadratic equations.

Week 3: Vectors and 2D relative motion

Level 1: Vectors

1. To be able to determine algebraic expressions for (v_x, v_y) for a vector defined in terms of a magnitude and a direction.
2. To be able to determine the x and y components of the vector sum of two vectors defined in terms of a magnitude and a direction.

3. To be able to identify a vector sum of two vectors displayed in a graphical representation.

Level 2: 2D Relative motion 1

1. To be able to apply the relative motion equation to create algebraic expressions for the x and y components of a velocity vector in frame A, given representations of that velocity vector in frame B and the velocity vector that relates the two frames.
2. To be able to determine quantities in frame A related to the velocity vector in frame A, given representations of that velocity vector in frame B and the velocity vector that relates the two frames.
3. To be able to determine qualitative relations between quantities in frame A related to the velocity vector in frame A, given representations of that velocity vector in frame B and the velocity vector that relates the two frames.

Level 3: 2D Relative motion 2

1. To be able to apply the relative motion equation to create algebraic expressions for the x and y components of a velocity vector in frame A, given representations of that velocity vector in frame B and the velocity vector that relates the two frames.
2. To be able to determine quantities in frame A related to the velocity vector in frame A, given representations of that velocity vector in frame B and the velocity vector that relates the two frames.
3. To be able to determine qualitative relations between quantities in frame A related to the velocity vector in frame A, given representations of that velocity vector in frame B and the velocity vector that relates the two frames.

Week 4: Projectile motion

Level 1: Projectiles 1

1. To be able to solve any well formulated projectile problem that does not require the solution of coupled equations or quadratic equations.
2. To be able to determine parameter changes for the change of one initial parameter.

Level 2: Projectiles 2

1. To be able to solve any well formulated projectile problem that requires the solution of coupled equations or quadratic equations.
2. To be able to determine qualitative relations between multiple parameters.

Level 3: Projectiles 3

1. To be able to answer qualitative questions that require knowing the relations between all parameters of projectile motion.

Week 5: Newton's second law

Level 1: Drawing FBDs

1. To be able to identify all forces acting on a body.
2. To be able to identify the correct free body diagram for a specific object.

Level 2: Equations from FBDs (Ramps)

1. To be able to write down Newton's Second Law equations in 2D, given a correct free body diagram for situations involving a block on a ramp.

Level 3: Equations from FBDs

1. To be able to write down Newton's Second Law equations in 2D, given a correct free body diagram for situations not involving a block on a ramp.

Level 4: Full problem

1. To be able to correctly identify free body diagrams and the resulting application of Newton's Second Law for situations involving one object.
2. To be able to solve the Newton's Second Law equations for unknown parameters.

Level 5: Newton's second law with two objects

1. To be able to correctly identify free body diagrams and the resulting application of Newton's Second Law for situations involving two objects connected by a taut string.
2. To be able to solve the Newton's Second Law equations for unknown parameters.

Week 6: Newton's first and third laws

Level 1: Identifying Newton's third law pairs

1. To be able to identify force pairs and to compare forces during a collision.

Level 2: Newton's third law (static)

1. To be able to apply Newton's Third Law to a situation in which forces are applied to a system of three blocks that result in zero acceleration and answer quantitative and qualitative questions.

Level 3: Newton's third law (accelerating)

1. To be able to identify Newton's Third Law pairs.
2. To be able to apply Newton's Third Law to a situation in which forces are applied to a system of three blocks that result in non-zero acceleration and answer quantitative and qualitative questions.

Level 4: Introduction to friction (as a constraint force)

1. To be able to apply Newton's Third Law to a situation in which forces are applied to a system of three blocks, with friction assuring that the blocks move as a group, that result in non-zero acceleration and answer quantitative and qualitative questions.

Week 7: Friction and uniform circular motion

Level 1: Static friction

1. To be able to determine direction and magnitude of static friction force between objects when they are or are not on the verge of moving relative to one another.

Level 2: Kinetic friction

1. To be able to determine direction and magnitude of kinetic friction force when there is relative motion between two objects.

Level 3: Static or kinetic friction?

1. To be able to determine whether force applied to a two-body system (with friction between the bodies) will cause the bodies to move together or slip relative to each other.

Level 4: Uniform circular motion

1. To be able to apply Newton's Second Law to situations involving circular motion in a horizontal plane which results in uniform circular motion.

Week 8: Universal gravitation and springs

Level 1: Gravitational forces

1. To be able to determine gravitational force between two point masses.
2. To be able to determine gravitational force between a spherical planet and a point mass.

Level 2: Orbital motion

1. To be able to determine forces and kinematic variables for planets (moons) moving in circular orbits around a central body (sun, planet).

Level 3: Springs

1. To be able to determine direction and magnitude of net force exerted on an object (including the force exerted by a spring) when that object is in equilibrium or at any position after released from a non-equilibrium position.

Appendix D

Algebraic Scaffolding Experiment Example Levels

Below are example versions of the two treatments in Level 5: a version with less scaffolding and a version with more scaffolding. The actual treatments had four versions in mastery-style delivery.

Less Scaffolding Treatment

All five questions refer to the situation below.

From the top of a tall cliff, two rocks are thrown downward with initial speeds v_1 and v_2. The initial speed of the second rock v_2 is four times the speed of the first rock v_1. For this problem, we will define the positive y-direction to be up with y=0 to be at the top of the cliff. Note that this definition will make all positions y and velocities v have negative values. Remember that the speed is defined as the magnitude of the velocity; so v_1 and v_2 are positive numbers, while all velocities in this problem will be negative. Therefore, for example, the initial velocity of rock 1 is equal to $-v_1$. We will also define t=0 to be the time when they are both thrown. The situation is shown in Figure D.1.

The equations you will use are the equations that describe the time dependence of position y and velocity v for an object moving with constant acceleration a:

$v = v_o + at$

$y = y_o + v_o t + 1/2at^2$

In this problem we are dealing with balls in free fall, so we know the value of the acceleration $a = -g = -9.81 m/s^2$, since we defined positive y to point up.

The end goal of this problem is to find the ratio of the rocks' velocities when the faster thrown rock has fallen twice the distance from the top of the cliff compared to the distance fallen by the slower rock.

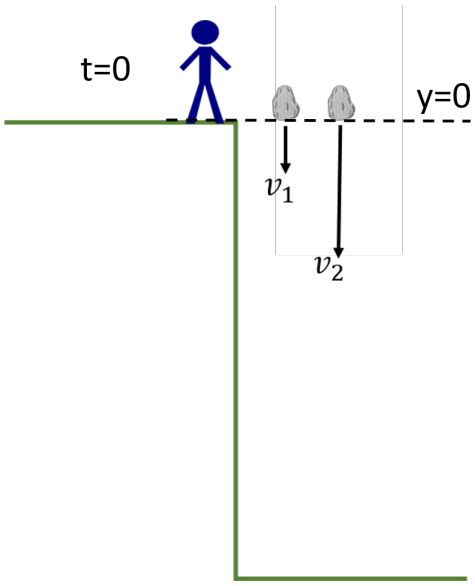


Figure D.1: Figure provided to illustrate situation in this version's problems.

The questions in this problem are structured to lead you to the final answer by asking for intermediate variables which can be obtained using the above two equations. Please work carefully as it is easy to make algebraic mistakes if you rush.

1) What is the correct way to express the relationship between v_1 and v_2, the speeds of rock 1 and rock 2, respectively?

a. $v_1 = 4v_2$
b. $v_1 = v_2$
c. $4v_1 = v_2$

2) The question wants to know the velocities of the rocks at a specific time t_f, the time when rock 2 has fallen twice as far as rock 1. Our first step is to determine an expression for t_f in terms of the initial speeds v_1 and v_2 and the acceleration due to gravity, g.

Which of the following expressions gives the correct expression for t_f, the time at which rock 2 has fallen twice as far as rock 1?

a. $t_f = 2\left(\frac{2v_2 - v_1}{g}\right)$
b. $t_f = 2\left(\frac{v_2 - 2v_1}{g}\right)$
c. $t_f = 2\left(\frac{v_1 - 2v_2}{g}\right)$
d. $t_f = 2\left(\frac{2v_1 - v_2}{g}\right)$
e. $t_f = 2\left(\frac{v_2 - v_1}{g}\right)$

3) What is $v_{1,f}$ the velocity of rock 1 at the time $t = t_f$, when rock 2 has fallen twice as far as rock 1, in terms of the speeds v_1 and v_2?

a.$v_{1,f} = v_1 - 4v_2$
b. $v_{1,f} = -5v_1 + 2v_2$
c. $v_{1,f} = -3v_1 + 4v_2$
d. $v_{1,f} = 3v_1 - 2v_2$
e. $v_{1,f} = v_1 - 2v_2$

4) What is $v_{2,f}$ the velocity of rock 2 at the time $t = t_f$, when rock 2 has fallen twice as far as rock 1, in terms of the speeds v_1 and v_2?

a.$v_{2,f} = 4v_1 - 3v_2$
b.$v_{2,f} = -4v_1 + v_2$
c. $v_{2,f} = 2v_1 - 5v_2$
d. $v_{2,f} = 2v_1 - v_2$
e. $v_{2,f} = -2v_1 - 5v_2$

5) Using the equation that relates the initial speeds v_1 and v_2, determine the ratio of the velocity of rock 2 to the velocity of rock 1 the time $t = t_f$, when rock 2 has fallen twice as far as rock 1.

a. $\left(\frac{v_{2,f}}{v_{1,f}}\right) = 5$
b. $\left(\frac{v_{2,f}}{v_{1,f}}\right) = 4$
c. $\left(\frac{v_{2,f}}{v_{1,f}}\right) = \frac{8}{3}$
d. $\left(\frac{v_{2,f}}{v_{1,f}}\right) = \frac{5}{3}$
e. $\left(\frac{v_{2,f}}{v_{1,f}}\right) = \frac{8}{5}$

More Scaffolding Treatment

All five questions refer to the situation below.

From the top of a tall cliff, two rocks are thrown downward with initial speeds v_1 and v_2. The second rock (v_2) is thrown with three times the speed of the first rock. For this problem, we will define t=0 to be the time when they are thrown and y=0 to be the top of the cliff, where they are thrown from. The situation is shown in Figure D.2.

The end goal of this problem is to find the ratio of the rocks' velocities when the faster thrown rock has fallen twice the distance from the top of the cliff compared to the distance fallen by the slower rock.

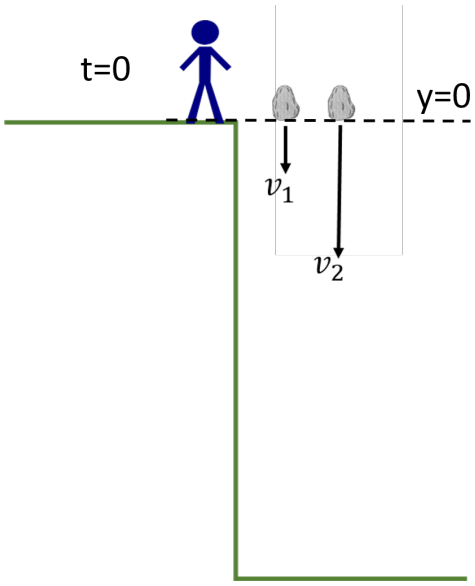


Figure D.2: Figure provided to illustrate situation in this version's problems.

This problem is algebra heavy, so the problem is going to step you through one way to solve the problem (though there are others), so you can practice manipulating and combining variables.

There will be a lot of variables to juggle in this problem, so be careful about labelling everything as belonging to rock 1 or rock 2. In the end, we'll want a ratio, which means things will cancel nicely if we can write everything in terms of one rock's variables. In this step-through, we'll use rock 1's variables, but this is not the only way to solve the problem.

(1) To begin with, we'll explicitly define the relationship between the initial speeds of the rocks. This is a piece of information we can use write variables in terms of each other.

(2) The question wants to know the rocks' velocities at a specific time, the time when rock 2 has fallen twice as far as rock 1. We need to find that time, and we can do it by expressing the condition that $2y_1 = y_2$ and solving for the time that makes that true. We can use the kinematics equation $y = y_o + v_o t + 1/2at^2$. Both are accelerating downwards due to gravity, and have original velocities downward, so if the positive y-direction is upwards, then we know:

$y_1 = -v_1 t - 1/2gt^2$

$y_2 = -v_2 t - 1/2gt^2$

Setting the condition that $2y_1 = y_2$, this becomes

$2\left(-v_1 t - 1/2gt^2\right) = \left(-v_2 t - 1/2gt^2\right)$

and you can solve for t, the time specified by the problem, in terms of v_1 and g.

(3 & 4) Once you have the time when the second rock has fallen twice as far as the first rock, you can use the kinematics equation $v = v_o + at$ to find each rock's velocity at the specified time. Remember that both

have initial negative velocities, and a negative acceleration due to gravity, $a = -g$. This turns the equations into:

$v_{1,f} = -v_1 - gt$

$v_{2,f} = -v_2 - gt$

(5) Once you have an expressions for $v_{1,f}$and $v_{2,f}$ in terms of v_1, you can up the ratio and cancel variables.

Refer back to this strategy map if necessary, but also try to understand how the parts work together.

1) What is the correct way to express the relationship between v_1 and v_2, the speeds of rock 1 and rock 2, respectively?

a. $v_1 = 3v_2$
b. $v_1 = v_2$
c. $3v_1 = v_2$

2) What is the time when rock 2 has fallen twice as far as rock 1?

a. $t = \frac{3}{2}\left(\frac{v_1}{g}\right)$
b. $t = 2\left(\frac{v_1}{g}\right)$
c. $t = \frac{5}{2}\left(\frac{v_1}{g}\right)$
d. $t = 3\left(\frac{v_1}{g}\right)$
e. $t = 6\left(\frac{v_1}{g}\right)$

3) What is the velocity of rock 1 at the time when rock 2 has fallen twice as far as rock 1?

a. $v_{1,f} = -\frac{4}{3}v_1$
b. $v_{1,f} = -\frac{3}{2}v_1$
c. $v_{1,f} = -2v_1$
d. $v_{1,f} = -3v_1$
e. $v_{1,f} = -4v_1$

4) What is the velocity of rock 2 at this time, in terms of v_1?

a. $v_{2,f} = -\frac{9}{4}v_1$

b. $v2, f = -3v_1$
c. $v_{2,f} = -4v_1$
d. $v_{2,f} = -\frac{9}{2}v_1$
e. $v_{2,f} = -5v_1$
f. $v_{2,f} = -10v_1$

5) Now, to find the ratio of their velocities, $\frac{v_{2,f}}{v_{1,f}}$.

a. $\left(\frac{v_{2,f}}{v_{1,f}}\right) = 1$
b. $\left(\frac{v_{2,f}}{v_{1,f}}\right) = \frac{5}{4}$
c. $\left(\frac{v_{2,f}}{v_{1,f}}\right) = \frac{3}{2}$
d. $\left(\frac{v_{2,f}}{v_{1,f}}\right) = \frac{5}{3}$
e. $\left(\frac{v_{2,f}}{v_{1,f}}\right) = 2$
f. $\left(\frac{v_{2,f}}{v_{1,f}}\right) = 3$
g. $\left(\frac{v_{2,f}}{v_{1,f}}\right) = 4$