

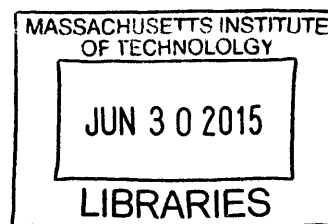
# Learning Experiments in a MOOC (Massive Open Online Course)

by

Christopher A. Chudzicki

B.A., Williams College (2010)

**ARCHIVES**



Submitted to the Department of Physics  
in partial fulfillment of the requirements for the degree of

Master of Science in Physics

at the


MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

© Massachusetts Institute of Technology 2015. All rights reserved.

**Signature redacted**

Author .....

  
Department of Physics  
January 16, 2015

**Signature redacted**

Certified by .....

David E. Pritchard  
Cecil and Ida Green Professor of Physics  
Thesis Supervisor

**Signature redacted**

Accepted by .....

  
Krishna Rajagopal  
Associate Department Head for Education



# Learning Experiments in a MOOC (Massive Open Online Course)

by

Christopher A. Chudzicki

Submitted to the Department of Physics  
on January 16, 2015, in partial fulfillment of the  
requirements for the degree of  
Master of Science in Physics

## Abstract

We present results from two treatment / control experiments in the 8.MReV: Mechanics ReView massive open online course (MOOC) run on edX.org during summer 2014. We compare the efficacy of physics homework problems: (1) traditional physics problems involving many skills, (2) deliberate-practice activities that train individual skills using the drag-and-drop format, and (3) analogous deliberate practice activities in multiple choice format. Using a common assessment, our results suggest that traditional instruction is more effective than deliberate practice activities cast in the multiple-choice format; comparison of traditional problems and drag-and-drop deliberate practice is so far inconclusive. Some evidence suggests users prefer the drag-and-drop format over multiple-choice and are more engaged in such problems.

In a separate experiment, we investigate the validity of the pre-test/post-test methodology in a MOOC environment where students receive feedback on the pre-test and can view the correct answer after finishing a pre-test problem. It seems that little learning occurs during the pre-test and that exposure to a problem on the pre-test usually does not provide students an advantage on the post-test.

Thesis Supervisor: David E. Pritchard  
Title: Cecil and Ida Green Professor of Physics



## Acknowledgments

I am deeply grateful for the support and mentoring of my advisor Professor David E. Pritchard and his postdoctoral assistant Zhongzhou Chen. I also thank all members of the RELATE group at MIT for their aid in statistical and edX-log-file analysis, especially Giora Alexandron for Zhou Qian for constructing student response and time-on-task matrices. I am grateful to Isaac L. Chuang for introducing me to physics education research, and to MITx and the MIT Physics Department for support during the Summer and Fall of 2014.



# Contents

<b>1</b>	<b>Learning Experiments in a MOOC</b>	<b>9</b>
<b>2</b>	<b>Anatomy of 8.MReVx: Mechanics ReView</b>	<b>11</b>
2.1	Background . . . . .	11
2.2	An edX Course . . . . .	12
2.2.1	Structure of a Generic Course . . . . .	12
2.2.2	Problem Nodes . . . . .	13
2.3	Structure of 8.MReVx . . . . .	14
2.3.1	Course Schedule . . . . .	14
2.3.2	Course Grading . . . . .	15
2.3.3	Discussion Forum . . . . .	16
2.3.4	Split-Test (“A/B”) Experiments in 8.MReVx2014 . . . . .	16
<b>3</b>	<b>Experiment: Deliberate Practice and a Comparison of Problem Formats</b>	<b>19</b>
3.1	Background . . . . .	20
3.1.1	The Nature of Expertise . . . . .	20
3.1.2	Training Expertise through Deliberate Practice . . . . .	21
3.2	Methods . . . . .	23
3.2.1	Participants . . . . .	23
3.2.2	Study Setup . . . . .	23
3.2.3	Treatments . . . . .	25
3.3	Efficacy of Deliberate Practice . . . . .	29

3.3.1	Results . . . . .	29
3.3.2	Discussion . . . . .	34
3.4	Comparison of Drag and Drop vs Multiple Choice Problem Formats .	36
3.4.1	Results . . . . .	36
3.4.2	Discussion . . . . .	41
3.5	Conclusions . . . . .	44
<b>4</b>	<b>Experiment: Does Pre-Test Feedback enhance Post-Test performance?</b>	<b>47</b>
4.1	Background . . . . .	47
4.2	Study Setup . . . . .	48
4.3	Results and Discussion . . . . .	49
4.4	Conclusions . . . . .	53
<b>A</b>	<b>Brief Description of the Atomic and Molecular Timing Algorithms</b>	<b>55</b>
A.1	Time interacting with problem (“atomic” time) . . . . .	55
A.2	Total time on problem and related resources (“molecular” time) . . . .	57
<b>B</b>	<b>Calculating Advantage Due to Exposure on Pre-test</b>	<b>59</b>



# Chapter 1

## Learning Experiments in a MOOC

In recent years massive open online courses (MOOCs) have emerged as a new model of education open to millions of students worldwide. Moreover, MOOCs offer researchers a wealth of data about the way that students learn through interactions with course material and each other. Much initial research into learning within MOOCs used patterns of resource usage to investigate broad questions in education about what student characteristics and behaviors correlate with learning [1, 2, 3].

In a complex field like education, the most reliable inferences about more fine-grained questions such as the efficacy of particular instruction techniques rely on experiments involving treatment and control groups to isolate effects. MOOCs are potentially a powerful environment in which to conduct education treatment/control (henceforth A/B) studies because of the large sample size and extensive user information collected. Moreover, MOOCs are natural place to study many questions related to problem format, interactivity, and computer-assisted learning. However, the analysis of A/B experiments in MOOCs is somewhat complicated by low completion rates and self-selection effects.

In this thesis we describe two A/B studies in the 2014 iteration of 8.MReVx: Mechanics ReView, run on the edX platform. The nature of the edX platform and the structure of 8.MReVx2014 in particular are described in Chapter 2.

Chapter 3 focuses on a study that explores two issues related to the design of more effective problems to build physics expertise: the deliberate practice of elementary

skills and the roll of interactive problem format. The solution of many standard problems in introductory classical mechanics requires the simultaneous execution of many skills and in such problems struggling students sometimes have difficulty identifying the source of their confusion. In studying expertise among world-class performers, Ericsson [4] identified *deliberate practice*—practice characterized by a singular focus on elementary skills, repetition, feedback, and the opportunity for improvement—as particularly important to the development of expertise. We investigate whether this deliberate practice framework, informed by research on expert–novice differences in physics, can be used within a MOOC to efficiently build problem-solving expertise by training specific skills. Additionally, we ask how the choice of problem format influences the effectiveness of our deliberate practice activities. In particular, we investigate how performance and learning compare when multiple-choice deliberate practice activities are replaced by informationally equivalent drag-and-drop problems, specifically designed to minimize extraneous cognitive load.

Chapter 4 describes a brief followup to an earlier study that measured learning in the 2013 iteration of 8.MReVx: Mechanics ReView using gain in score from a pre-test to a post-test [3]. Unlike in traditional pre-test/post-test settings, users in a MOOC receive feedback during the pre-test and are able to view the correct answers after finishing each problem. We present preliminary results suggesting that for most problems this difference is unimportant.

As with any large endeavor, running a massive open online course and analyzing the gathered data is a group effort. My role in this work was twofold: I helped to develop many of the deliberate practice activities discussed in Chapter 3 and was also primarily responsible for analyzing student data relevant to the experiments discussed here, once response and time-on-task matrices had been constructed by other team members.

# Chapter 2

## Anatomy of 8.MReVx: Mechanics ReView

### 2.1 Background

The online course 8.MReVx: Mechanics ReView began as a 3-week residential course offered between the Fall and Spring semesters to MIT students who struggled in the Fall semester required course *8.01: Classical Mechanics*. The residential version, 8.MReV: Mechanics ReView, was offered for the first time in January 2009 and taught using flipped-classroom methods. During the summer of 2012 8.MReV: Mechanics ReView was offered as a free, massive open online course (MOOC) on the LON-CAPA platform. Since then, *Mechanics ReView* has been offered as the MOOC 8.MReVx on the edX during Summer 2013 and Summer 2014.

The Summer 2014 iteration of 8.MReVx was one of the first edX MOOCs to use the split-test functionality to implement controlled content experiments. In this chapter we describe the structure of an edX course with particular focus on the 2014 iteration of 8.MReVx: Mechanics ReView and the implementation of split-test content experiments within edX.

## 2.2 An edX Course

### 2.2.1 Structure of a Generic Course

The vast majority of edX course content is viewed by students within a “Courseware Tab”. The structure of edX Courseware is depicted in Figure 2-1 and largely mirrors the hierarchical structure of a traditional textbook:

- At top level, Courseware contains several **chapters**. Each chapter represents material for a given time period. Often, courses contain one chapter per week.
- Each chapter can contain several **sequentials**, analogous to sections in a textbook.
- Each sequential contains several **verticals**, analogous to pages in a textbook.
- Each vertical can contain an unlimited number of **nodes**. Chapters, sequentials, and verticals are purely organization structures that contain content. Nodes *are* the content, and can be of a variety of types e.g., HTML nodes (for text, examples, or figures), video nodes, or interactive problem nodes.

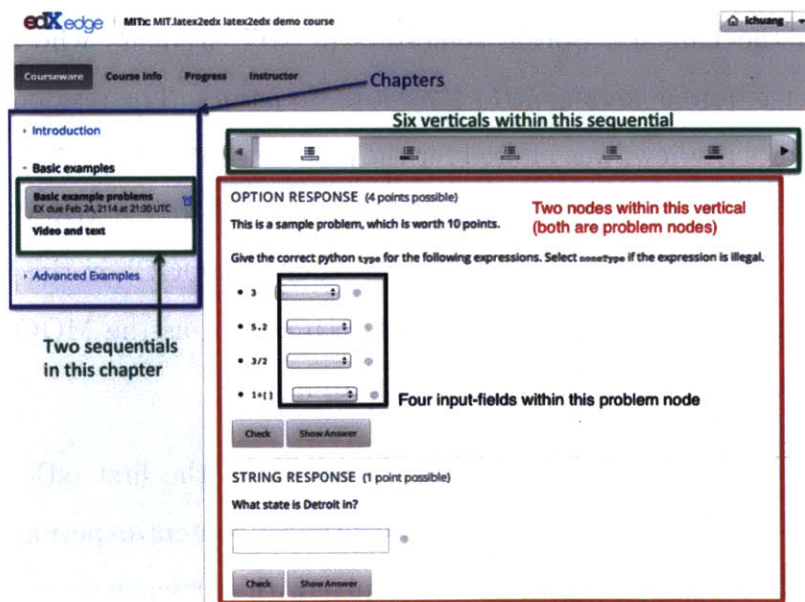


Figure 2-1: The hierarchical edX course structure

Content becomes visible to students when it is “released”. Release and due dates are set sequential by sequential, though it is common to release all sequentials within a chapter simultaneously.

In addition to the Courseware tab that contains course content, edX courses contain other tabs such as Progress, Course Info, and Discussion. The course discussion forums can be accessed either by the Discussion tab or by the discussion nodes embedded within verticals.

## 2.2.2 Problem Nodes

Although edX courses are structured much like a traditional textbook, their content can be much more interactive. Each problem node has a **check** button that provides immediate feedback to users as to whether their submission was correct or incorrect. (Additionally, it is possible to provide more nuanced feedback to address particular errors). Each problem node can be assigned a particular number of attempts (possibly infinite), corresponding to the number of times a student is allowed to hit the **check** button for that problem.

Problem nodes can be constructed in several formats ranging from multiple choice to symbolic input to drag-and-drop activities. Each problem node can contain several parts termed **input-fields**. The first problem node in Figure 2-1 contains four input-fields; the second contains one input-field. When the user submits an answer with the **check** button, all input-fields are graded simultaneously. Input-fields are usually graded independently from one another and correct/incorrect feedback is given on each input field.

Problem nodes are the only component of an edX course that can be graded. Each problem can be assigned a number of points (possibly fractional). When graded, problem node points are split evenly among input-fields. For the purpose of calculating a total course grade, each sequential that contains graded problems can then be assigned to differently weighted categories (e.g., homework, quiz).

## 2.3 Structure of 8.MReVx

### 2.3.1 Course Schedule

The 2014 iteration of 8.MReVx: Mechanics ReView contains 12 weeks of graded material and 2 weeks of ungraded, optional material. In most cases, material for each week corresponds to three chapters within the course—one chapter for instruction material, one chapter for that week’s homework, and one chapter for a weekly quiz<sup>1</sup> (see Table 2.1). The material for a given week was released three weeks before it was due, with the quiz and homework for a given week due simultaneously.

To these general rules there are a few exceptions in which the homework, quiz, or both for two consecutive weeks is combined (e.g., the first two weeks of the course). Moreover, the first six weeks of the course were released simultaneously.

The relatively large 3-week delay between release dates and due dates was intended to allow users flexibility in their summer schedules. This seemed especially important because 8.MReVx runs during the summer when long vacations are common.

Table 2.1: A partial course outline for 8.MReVx: Mechanics ReView-2014

	Chapter	Grading Category	Release Date	Due Date
Week 1–2	1: Newton’s Laws of Motion	Checkpoints	Thursday, 5/29/2014	3 weeks+3 days later
Week 1–2	2: Interactions and Forces	Checkpoints	Thursday, 5/29/2014	3 weeks+3 days later
Week 1–2	Homework for Weeks 1–2	Homework	Thursday, 5/29/2014	3 weeks+3 days later
Week 1–2	Quiz for Weeks 1–2	Quiz	Thursday, 5/29/2014	3 weeks+3 days later
	⋮			
Week 7	7: Linear Momentum	Checkpoints	Sunday, 7/20/2014	3 weeks later
Week 7	Homework for Week 7	Homework	Sunday, 7/20/2014	3 weeks later
Week 7	Quiz for Week 7	Quiz	Sunday, 7/20/2014	3 weeks later
Week 8	8: Mechanical Energy	Checkpoints	Sunday, 7/27/2014	3 weeks later
Week 8	Homework for Week 8	Homework	Sunday, 7/27/2014	3 weeks later
Week 8	Quiz for Week 8	Quiz	Sunday, 7/27/2014	3 weeks later

<sup>1</sup>Why separate each week’s material into three chapters—why not put all sequentials for the same week within one chapter? 8.MReVx2014 inherited this structure from 8.MReVx2013, in which quiz due dates were delayed from the corresponding homework due date by one week. At the time, this was most easily achieved within the edX platform by separating instruction material, homework, and quiz into separate chapters.

## 2.3.2 Course Grading

In order to “pass” the course and receive an edX Honor Code Certificate, students in 8.MReVx: Mechanics ReView needed to earn a grade of at least 60%. Graded problem nodes fell into one of five categories:

- **Checkpoints**, 7% of total grade: embedded within instructional material. Five sequentials (roughly one-tenth of all checkpoints) worth of checkpoints were dropped.
- **Homework Problems**, 34% of total grade: there was roughly one homework assignment per week, as described above.
- **Quiz Problems**, 36% of grade: there was roughly one quiz assignment per week, as described above.
- **Midterm Exam**, 7% of the grade: A cumulative midterm exam was given in lieu of a week-9 quiz.
- **Final Exam**, 16% of total grade: A cumulative final exam was given due one week after the last graded instructional chapter. (The final exam was also given before week 1 as an ungraded pre-test. The pre-test was hidden from students after its due date.)

Table 2.2 shows the number of input fields and problem nodes per category.

Table 2.2: Number of input fields and problem nodes per grading category <sup>2</sup>

	Checkpoint	Homework	Quiz	Midterm	Final
Input Field	462	350	107	13	23
Problem Node	226	253	52	13	17

Most graded problems allowed students multiple attempts. Checkpoint and homework problems frequently allowed students up to 10 attempts, except when this would

---

<sup>2</sup>Because of split-test experiments, different users are exposed to a slightly different number of problems. These numbers are averages over split-test experiments.

significantly increase the chance of correctly answering by guessing. For example, a five-choice multiple-choice Checkpoint or Homework problem might be given 2 or 3 attempts, but a numerical response Checkpoint or Homework problem would usually be given 10 attempts. In contrast, quiz problems and Pre-Test/Midterm/Final-Exam problems were generally given 3 or fewer attempts.

### 2.3.3 Discussion Forum

8.MReVx: Mechanics ReView uses the edX Discussion feature to promote interaction between students. All students can make posts to the discussion forum, and were encouraged to ask general questions about physics or specific questions about physics problems within the course. However, students were explicitly told that posting final answers to graded course problems before their due dates was not permitted.

8.MReVx2014 benefitted from the hard work of 87 dedicated “Community TAs”, the vast majority of whom were recruited from the pool of certificate-earners in 8.MReVx2013. Many community TAs were very active in 8.MReVx2014, answering questions and policing the discussion forums. In the few instances when the final answer to a graded problem was posted before a due date, it seemed that community TAs quickly edited the original post to remove this information.

### 2.3.4 Split-Test (“A/B”) Experiments in 8.MReVx2014

Usually, all students in an edX course see exactly the same content. In Spring 2014, edX released the ability for courses to implement **split-tests**. In a single split-test, the entire user population is partitioned into two or more groups and each group is given separate course material. The split-test process is illustrated in Figure 2-2. 8.MReVx: Mechanics ReView-2014 used split-tests to implement seven separate learning experiments distributed throughout the course.

Several aspects of edX split tests are important to note:

- A split test can contain any type of edX content: html pages, problems, videos, etc, and can be graded or ungraded

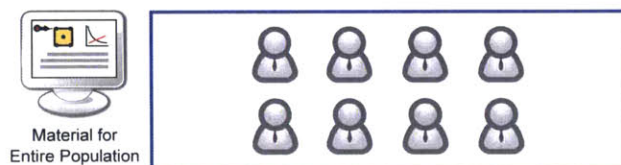


- A single course can contain any number of split tests.
- A single split test can contain two or more groups.
- All users participate in all split tests. Users cannot opt-out of a split test without opting-out of the course itself.
- Users assignment to groups within a split-test is random.
- Split tests can be implemented at either the sequential level (so that group A and group B see different sequentials within the same chapter) or vertical level (so that group A and group B see different verticals within the same sequential).
- Split tests are stable throughout the course—that is, the same partitioning of users can be used in different parts of the course. For example, Partition 1 could be used the first week, partition 2 the third week, and partition 1 could be used again during the tenth week.

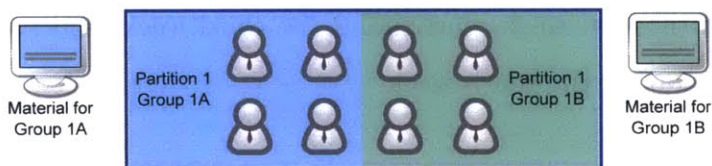
### **Student Awareness**

The About Page on edx.org for 8.MReVx2014 where students enroll in the course explicitly informed students that the course contained split-test experiments and that not all students would receive the same material. Additionally, some students became aware of specific split-test experiments through the discussion forums because although the A group and B groups in a two-group “A/B” experiment have separate courseware material, the discussion forum is *not* separated by group.

### Normal Course Material (Not a Split Test)



### Split Test: Partition 1



### A Different Split Test: Partition 2

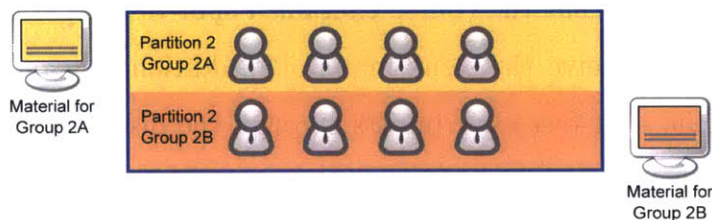


Figure 2-2: Schematic of an edX course containing common material and two separate split-tests

## Chapter 3

# Experiment: Deliberate Practice and a Comparison of Problem Formats

Over the past several decades there has been extensive research exploring the nature of expertise and the differences between experts and novices in physics, chess, athletics, music, and other fields. Initial work by Ericsson in the 1990s [4] showed that in many fields the transition from novice to expert has more to do with practice than innate talent. However, not all practice leads to expertise. *Deliberate practice*—characterized by a singular focus on elementary skills, repetition, immediate feedback, opportunities for improvement, and self-reflection—was identified as a specific type of practice especially important to the development of expertise. *Cognitive load theory* [5] suggests that deliberate practice can be enhanced by reducing *extraneous* cognitive load, freeing the learner’s working memory to focus on the salient aspects of deliberate practice activities rather than irrelevant details such as those related to problem format.

In this chapter we describe our most ambitious split-test experiment in 8.MReVx: Mechanics ReView 2014. This experiment asks two separate but related questions. First, we investigate whether the deliberate practice framework, informed by research on expert–novice differences, can be used to design activities that efficiently build physics expertise in introductory mechanics students. Second, we ask how the choice of problem format affects learning in our MOOC. In particular, we investigate how perfor-

mance and learning compare when multiple-choice deliberate practice activities are replaced by informationally equivalent drag-and-drop problems, specifically designed to minimize extraneous cognitive load.

We begin by discussion previous work on the nature and development of expertise, then describe the setup of the deliberate practice study undertaken in Units 10, 11 and 12 of 8.MReVx2014. We present and discuss results about problem content (deliberate practice vs traditional problems) separately from problem format (multiple choice vs drag-and-drop).

## 3.1 Background

### 3.1.1 The Nature of Expertise

The first step in designing material to build physics expertise is identifying differences between experts and novices. Research on expertise has revealed important differences between the way that novices and experts approach problems and organize their knowledge.

- Experts quickly identify important information and this information triggers meaningful inferences.

One way experts quickly identify important features is by “chunking” pieces into meaningful patterns. Chess masters can quickly memorize board positions by chunking pieces into strategically relevant clusters [6] and electrical engineers chunk components in circuit diagrams into functional substructures [7]. Some work [8] suggests that in physics, novices have more difficulty using key features to trigger relevant inferences than identifying the features themselves.

- Experts organize knowledge around important ideas and principles.

Experts in physics have been shown to categorize problems [9] and plan solutions [8] by underlying principle (e.g., conservation of momentum) whereas novices categorize problems and plan by surface features (e.g., the problem contains a spring).

- Expert’s knowledge is conditionalized—experts know in what situations their knowledge and tools are valid.
- Experts exhibit fluent retrieval of their knowledge.

### 3.1.2 Training Expertise through Deliberate Practice

Consider the following introductory mechanics problem:

(“Kick the stick”) A stick of length  $\ell$  is at rest along the  $y$ -axis and is suddenly kicked at one end in the  $x$ -direction. Immediately after the stick is kicked, its center-of-mass moves rightward at speed  $v$ . At what angular speed  $\omega$  does the stick rotate after being kicked?

An expert might apply conservation of angular momentum to quickly solve this problem. Table 3.1 lists the numerous elementary skills, many of which an expert may perform automatically, that are needed for a correct solution.

- 
1. parsing the question and perhaps generating a pictorial representation
  2. identifying the relevant physical model (e.g., conservation of angular momentum),
  3. selecting a useful reference point for angular momentum,
  4. identifying the stick’s moment of inertia
  5. decomposing a rigid body’s angular momentum in spin and orbital contributions
  6. selecting a useful expression for each contribution to angular momentum (e.g.,  $\vec{L} = \vec{r} \times m\vec{v}$ ,  $L = r_{\perp}mv$ ,  $L = rmv_{\perp}$ , or  $L = I\omega$ ) and identifying the relevant variables,
  7. (depending on the approach in Step 6) evaluating a cross product.
- 

Table 3.1: Elementary skills needed to solve the “Kick the stick” problem.

This is an example of a problem that might be assigned in a standard introductory mechanics class to emphasize the crucial role that picking an appropriate reference point plays in applying the conservation of angular momentum. Throughout a traditional course, physics students are primarily trained by being assigned many such problems relating to various topics. These problems are “full-game” problems in that

they require the coordinated execution of many skills, as in a game of chess or an athletics competition. Physics students are sometimes given simpler practice problems that ostensibly focus on a single skill—for example, a problem asking students to calculate the angular momentum of a rigid body. Often times, such a task may appear as a single chunk to an expert, but actually involves multiple steps (e.g., steps 5–7 above).

Because the above problem requires so many elementary skills, struggling students may not be able to identify the source of their confusion if they are told their answer is incorrect. Even students who do successfully complete the problem may not be able to identify the key features of their solution among so many steps. Ericsson’s original work on the development of expertise [4] and subsequent studies [10, 11, 12] suggest that practicing primarily through such “full game”, real-world scenarios is not an effective way to build expertise. Instead, novices should supplement full games with *deliberate practice*. Deliberate practice activities (DPAs) have several key features.

- Deliberate practice activities focus on one or two elementary skills.
  - This allows students and instructors to quickly identify deficiencies and competencies and may make it easier for students to identify the “take-away message” of each problem.
- Deliberate practice activities have short duration, high repetition, and provide immediate feedback.
  - Short duration provides more opportunities for feedback in a fixed practice time. Immediate feedback and high repetition provides an opportunity for learning, allowing students to make adjustments between problem attempts.

In our study, we use the idea of deliberate practice to develop short physics problems each of which is designed to target one or two elementary skills, including many of those in Table 3.1.

Our effort to develop deliberate practice activities to build expertise in physics, a highly cognitive, open-ended domain, differs from previous work in two fundamental ways. First, the deliberate practice framework was developed mostly in music, typing, and athletics, all domains with a heavy emphasis on kinesthetic skills (see [13, 14, 15] for reviews), though chess, a highly cognitive skill, has also been studied [4, 11]. More recent work has examined the role of deliberate practice in other cognitive areas including professional writing [16] and medicine [17, 18]. The only study of which we are aware that advocates deliberate practice in physics education [19] focused much more heavily on interactive lectures and peer-instruction than the principles of deliberate practice. Second, previous research has tended to use deliberate practice only as a lens through which to view existing material in order to gain insight on the development of expertise, not for the construction of instructional tools. Gifford’s Doctor–Coach pedagogy [17] is a notable exception to this trend.

## **3.2 Methods**

### **3.2.1 Participants**

The participants were users in the MOOC 8.MReVx: Mechanics ReView 2014. Of the roughly 15,000 users who signed up for 8.MReVx2014, only 614 ever interacted with this particular experiment. Not all of these 614 users interacted significantly with the experimental content, and hence not all will be included in our analysis—the conditions for inclusion in analysis will be discussed in more detail below.

### **3.2.2 Study Setup**

Figure 3-1 depicts the structure of our study, which takes place during the last three graded units of our course (Unit 10: Rotation & Translation; Unit 11: Angular Momentum; Unit 12: Gravity and Orbits). All students in our course are randomly split into three groups (A, B, or C). The split into groups is stable over the course of our study. During each unit (10, 11, and 12) all students receive homework in two

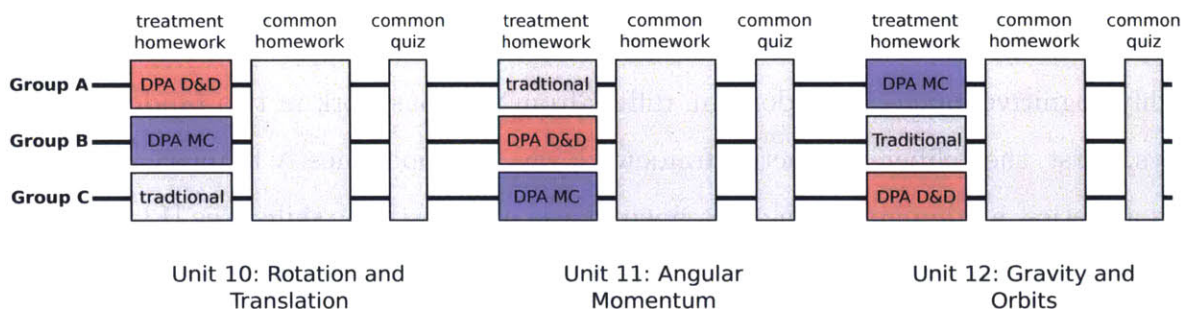


Figure 3-1: Each group receives homework in two problem sets, including common problems and treatment problems, described in Section 3.2.3. Learning in each group is evaluated through common end-of-unit quizzes testing transfer from common material. Treatment assignment is rotated during the study.

problem sets:

- The first set contains the treatment, which varies from group to group. The treatment consists either of traditional homework problems (control) or one of two variations of deliberate practice activities that differ in format (multiple choice, MC, vs drag-and-drop, DD).
- The second set contains traditional, “full-game” homework problems common to each group.

In order to evaluate the efficacy of deliberate practice, each unit culminates in a quiz that is common to all groups. The quiz consists of traditional problems only (conceptual and numeric/symbolic response). Quiz problems are designed to test transfer from material common to each of the three group homeworks. Some (not all) quiz and common homework problems have been used in previous iterations of the 8.MReVxMOOC.

In order to treat all participants in our study equally, the treatment assigned to each group rotates between each unit. Thus, Group C received the traditional control problems in Unit 10, deliberate practice problems in the drag-and-drop format during Unit 11, and deliberate practice problems in the multiple-choice format during Unit 12.



### 3.2.3 Treatments

Our study uses three treatments that differ in problem design (traditional vs. deliberate practice activities, DPA) and problem format (multiple-choice deliberate practice, MC vs. drag-and-drop deliberate practice, DD).

#### Traditional Treatment

The traditional problem sets are a mix of conceptual and full-game problems implemented through the multiple-choice, checkbox, drop-down, numerical input, and symbolic math input problem formats. The full-game problems vary in complexity from single-principle (e.g., the problem uses conservation of energy only) to multi-principle (several conservation laws or net force laws are used). Some of these problems are broken into parts, others are not. These problems mirror homework in a standard physics course, in previous iterations of the 8.MReVx MOOC, and are similar in style to the common problems and quiz problems given to all groups in Units 10, 11, and 12.

#### Deliberate Practice Treatments

The deliberate practice treatment consists of short problems targeting specific elementary skills used in solving full-game problems. Table 3.2 gives some examples of the elementary skills our problems are designed to target. (We believe these skills are useful to students on the common follow-up assessment, though the assessment requires additional skills as well.) Because the deliberate practice problems take a relatively short amount of time by their very nature, we are able to include several problems (usually at least 4) on each specific skill. Feedback and the opportunity to improve is a crucial part of deliberate practice and for this reason we wrote thorough solutions (many containing useful figures) to all deliberate practice problems. Solutions are viewable after a student has either successfully completed the problem or has been marked incorrect on all available attempts.

Table 3.3 shows the number of problems on the deliberate practice and tradi-

Unit	Skill
10	basic application of Newton's 2nd Law in linear and rotational form
10	identifying the type and direction of frictional forces that accelerate rolling objects
10	identifying the mathematical relationships imposed by physical constraints (e.g., <i>rolling without slipping</i> or <i>ideal rope</i> or <i>rope does not slip</i> )
11	identifying the appropriate quantities to be used for $r$ , $r_{\perp}$ , $v_{\perp}$ , and $\theta$ in angular momentum expressions $L = rmv \sin \theta = r_{\perp}mv = rmv_{\perp}$
11	identifying angular momentum reference points with specific properties
12	relating gravitational potential energy graphs and physical situations
12	identifying forces that do work and provide torque to change energy and angular momentum
10, 11, 12	identifying situations in which various conservation laws apply and why (Finding Error Problems)

Table 3.2: Some elementary skills targeted by deliberate practice activities

Table 3.3: Number of physics problems on treatment homework

	Unit 10	Unit 11	Unit 12
Deliberate Practice (DD)	21	19	19
Deliberate Practice (MC)	21	19	19
Traditional	6	10	6

tional homework treatments. Although the deliberate practice treatment contains many more problems than the traditional treatment, both treatments cover the same material and are intended to take roughly the same amount of time. To help isolate the effect of deliberate practice, we also wrote solutions to all problems used in the traditional treatment<sup>1</sup>

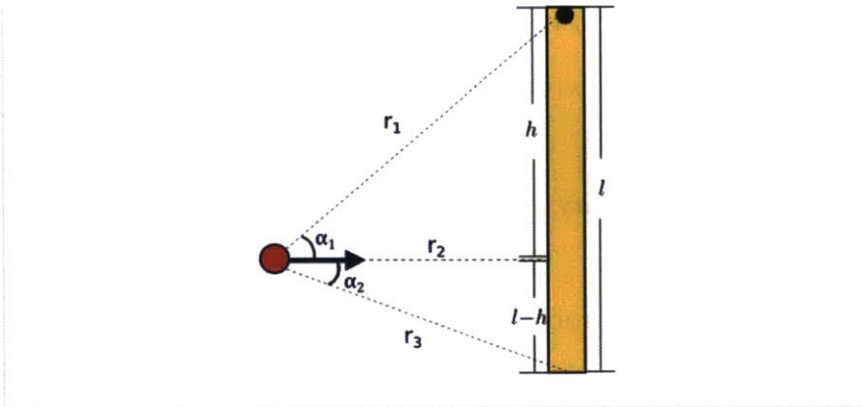
One type of deliberate practice activity—Finding Error problems—were used in all three units and deserve explanation. These activities present students with a physics problem and a solution to this problem that may or may not contain an error. Students are asked to decide whether or not the solution does in fact contain an error, and if it does, students must indicate what is the error and what information in the problem statement indicates that this is an error. Like worked example and problem completion tasks [5, 20], these Finding Error problems aim to help students focus on especially important aspects of problem by eliminating tasks such as algebraic manipulation that are only tangentially related to the physics and contribute to extraneous cognitive load.

### **Multiple Choice vs. Drag-and-Drop**

In order to investigate the influence of problem format on the effectiveness of deliberate practice, we created two versions of the deliberate practice treatment. One version (MC) administers deliberate practice activities through the standard multiple-choice (MC), drop-down, and checkbox problem formats. The second version (DD) uses edX’s drag-and-drop (DD) problem format exclusively. The drag-and-drop problem format is a very flexible problem format in which users drag objects onto a target image to indicate the answer to a question. It was hypothesized that the multiple-choice format might cause users to split their attention between different parts of the page (the choices and the problem statement) thereby increasing extraneous cognitive load through the “split-attention effect” [5, 21]. In contrast, drag-and-drop problems tend to collocate the problem statement and the figure, reducing cognitive load.

---

<sup>1</sup>Common problems do not have staff-created solutions. Student-created solutions may exist in the course wiki.



What is  $r$ ?

- $r_1$
- $r_2$
- $r_3$
- 0

What is  $\alpha$ ?

- $\alpha_1$
- $\alpha_2$
- 0

(a) Student Task: select appropriate values of  $r$  and  $\alpha$  to be used in  $L = rmv \sin \alpha$  by marking multiple-choice radio buttons.

(b) Student Task: select appropriate values of  $r$  and  $\alpha$  to be used in  $L = rmv \sin \alpha$  by dragging indicator to figure.

Figure 3-2: Comparison of multiple-choice and drag-and-drop format for a variable identification deliberate practice activity.

The deliberate practice activities in each treatment are as similar as possible within the constraints imposed by the different formats. Figure 3-9 shows two versions of the same deliberate practice activity, one in each format. The problem asks users to indicate what quantities should be used as the angle  $\alpha$  and the distance  $r$  in the expression  $L = rmv \sin \alpha$ .

There is an important difference between the two formats that could potentially confound comparison of results between formats when users are given multiple attempts at an activity, as they are in all cases. The drag-and-drop activity corresponds to a single edX input-field (users indicate  $r$  and  $\alpha$  simultaneously) whereas the multiple choice version corresponds to two separate input fields (users indicate  $r$  and  $\alpha$  separately). As a consequence, when users hit **check**, the multiple choice version indicates whether the value of  $r$  is correct/incorrect and whether the value of  $\alpha$  is correct/incorrect, whereas the drag-and-drop version only indicates whether the pair  $(r, \alpha)$  is correct/incorrect. That is, *the multiple choice group receives more feedback per attempt.*

## 3.3 Efficacy of Deliberate Practice

### 3.3.1 Results

Of the 614 users who accessed at least some of the material in this split test experiment, not all of them accessed a significant fraction of the material. Figure 3-3 depicts attempt rates for the treatment homework and common quiz in Units 10, 11, and 12. In all three units a large group of students attempted all of the treatment homework and all of the quiz (upper right corner) and a large group of students attempted none of the homework and none of the quiz (lower right corner). The relatively small off-diagonal values in these completion heat maps indicate a high correlation between attempting the homework and attempting the quiz.

In order to guarantee that the users we consider interacted with the treatment homework to a significant extent, we restrict our attention in the remainder of this

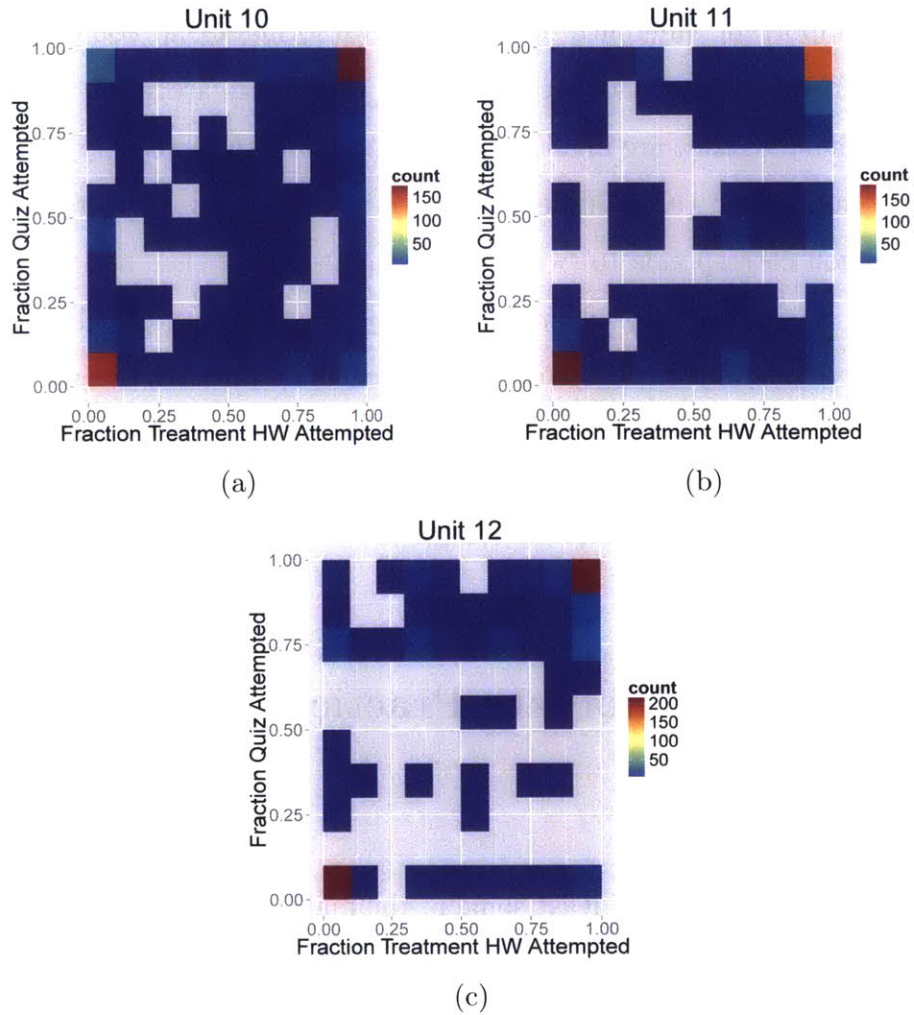


Figure 3-3: Treatment Homework and Quiz joint completion rates for Units 10, 11, and 12. A total of (219,205,280) during units (10,11,12) students completed at least 70% of the treatment homework and at least 70% of the common quiz.

section to only those who completed at least 70% of the treatment homework *and* at least 70% of the common quiz. This cut-off leaves a total of 219 students for Unit 10, 205 students for Unit 11, and 280 students for Unit 12. Note that post-selection *does not* guarantee that users completed the common quiz assessment before they completed the treatment homework.

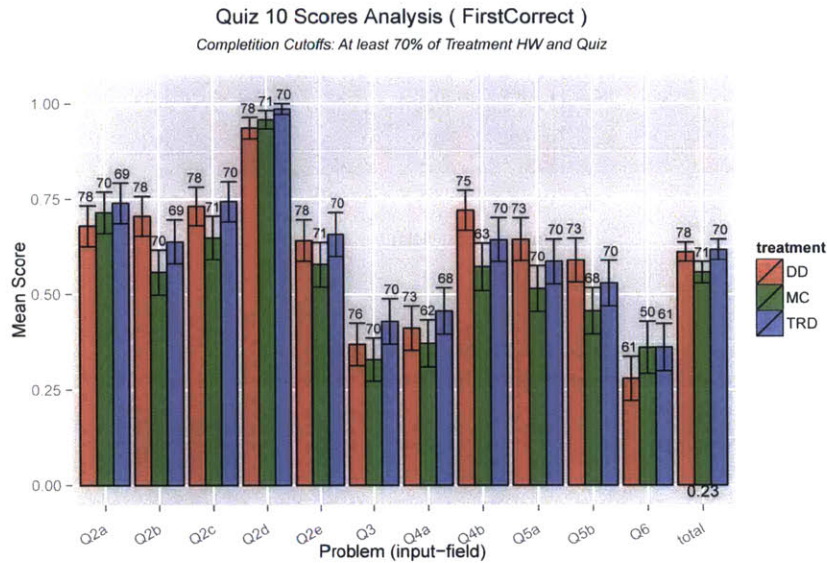


Figure 3-4: Comparison of first-attempt quiz scores for the three treatment groups: traditional instruction (TRD), deliberate practice in the drag-and-drop format (DD), and deliberate practice in the multiple-choice format (MC). Bar height indicates the mean score for that group. Error bars show one standard error in the mean. The number above each bar indicates the number of users from each group who attempted that quiz question. Total quiz score is calculated per user with unanswered questions ignored.



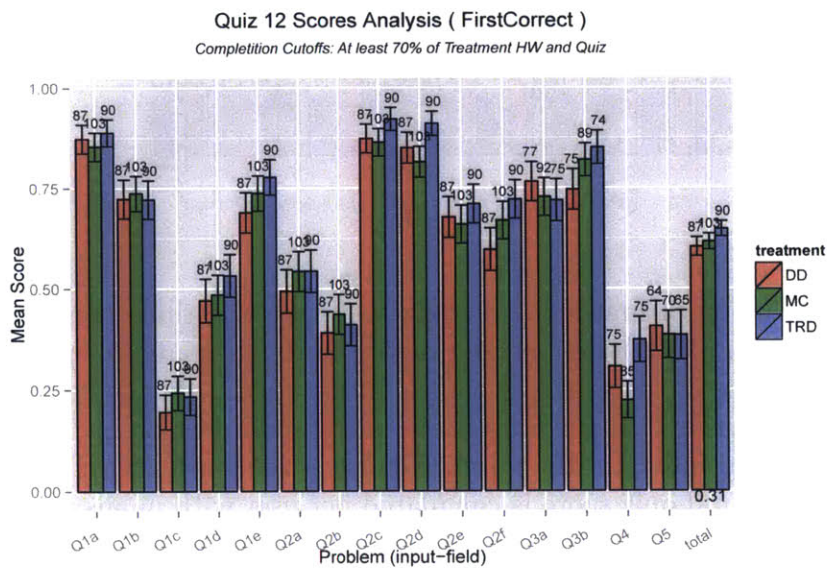
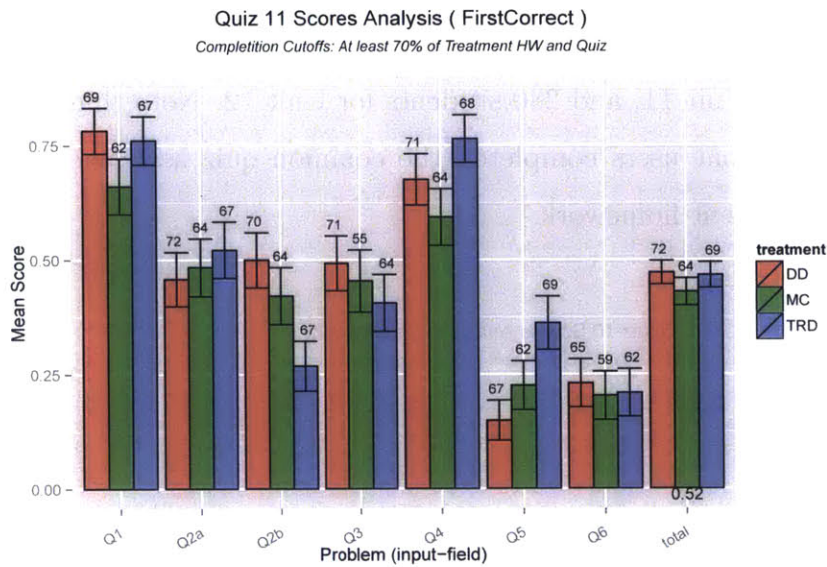


Figure 3-4: *continued*

The performance of each treatment group on common quizzes is summarized in Figure 3-4 and Table 3.4. Generally, users were permitted about 3 attempts on quiz questions in 8.MReVx2014. These scores are based only on first attempt, which shows the largest differences between groups. Using first-attempt-correct as the correctness criterion also helps ensure that the total scores are at least somewhat normally dis-



treatment	Total Score, Quiz 10				Total Score, Quiz 11				Total Score, Quiz 12			
	group	<i>N</i>	mean	sd	group	<i>N</i>	mean	sd	group	<i>N</i>	mean	sd
Delib. Prac. (DD)	A	78	0.61	0.23	C	72	0.47	0.22	B	87	0.59	0.22
Delib. Prac. (MC)	B	71	0.56	0.24	A	64	0.43	0.24	C	103	0.61	0.20
Traditional (TRD)	C	70	0.62	0.23	B	69	0.47	0.23	A	90	0.65	0.18
<i>p</i> (ANOVA)	0.23				0.52				0.22			

Table 3.4: Comparison of the Total Quiz Score, calculated per person based on first-attempt correct rates, for each of the three treatment groups on Units 10, 11, and 12. Standard one-way ANOVA was used to calculate *p*-values to test whether all three means are equal.

Table 3.5: Differences in total quiz score, averaged over quizzes 10, 11, and 12

comparison	difference in means	sd	<i>Z</i> -score	<i>p</i> -value
DD – MC	0.023	0.021	1.11	0.27
MC – TRD	–0.047	0.021	–2.22	0.026
TRD – DD	0.023	0.021	1.14	0.26

tributed and not skewed too far toward the upper limit of 1. Because users needed to earn only 60% of the points available in our course to earn a certificate, we did not penalize students for skipping questions when calculating the per-user quiz total score. That is, total score for each user on a given quiz was computed by

$$\text{total score for user} = \frac{\text{number of quiz questions answered correctly}}{\text{number of quiz questions attempted}}. \quad (3.1)$$

(Recall, however, that only those users who completed at least 70% of the treatment and quiz are included in this analysis, so relatively few scores were dropped.) Per-user total scores are independent from one another and can be analyzed using standard statistical methods. Using a one-way ANOVA, we did not detect a significant difference in the mean total scores for any particular chapter. However, on all three quizzes the group receiving traditional instruction did out-perform the group receiving deliberate practice problems in the multiple choice format, having a higher total quiz score 4.7% averaged over the three units (see Table 3.5). Using a *t*-test, this difference was

significant at  $p = 0.027$ ; no other cumulative differences were significant.

Although there are no obvious differences in quiz score, it is conceivable that different treatment groups spent different amounts of time-on-task for solving each quiz problem. To investigate this possibility, we analyzed edX log files to determine the time per problem for each user. For this analysis, we used the "molecular time" algorithm as our operational definition of time-on-task, which is intended to include time spent viewing a problem and time spent on related resources when estimating time-on-task. Roughly speaking, the molecular time algorithm includes all time between opening a problem and finishing a problem, including time spent on other tasks between these two events, as long as time spent on other tasks is not too large. For a detailed description of this algorithm, see Appendix A.

Figure 3-5 shows the time-on-task (molecular timing algorithm) for each problem node in the common quiz for Unit 11. Because timing data is highly skewed, we used the nonparametric Mann-Whitney  $U$ -test to perform pairwise comparison between groups of time-on-task during the quiz. For continuous random variables such as time-on-task, the Mann-Whitney  $U$ -test can be interpreted as testing for a difference in median [22]. No between-group differences in *total time* spent on quiz are statistically significant at  $p < 0.1$ . (Two differences between groups in time spent on specific problems were significant, both at  $p = 0.04$ . However, since there are 18 quiz problems, 3 groups, and 54 pairwise comparisons in total, it is not at all unlikely that these two differences occurred by chance).

### 3.3.2 Discussion

Despite receiving (and completely at least 70% of) markedly different treatment homework as preparation for the common quizzes, the deliberate practice and traditional instruction groups took similar amounts of time to complete the common quizzes in units 10, 11, and 12, and no significant difference in quiz score was observed between groups for a given unit. When all three quizzes are considered, the data suggest that the traditional treatment outperforms the multiple-choice deliberate practice treatment, but no meaningful conclusions about the efficacy of drag-and-drop

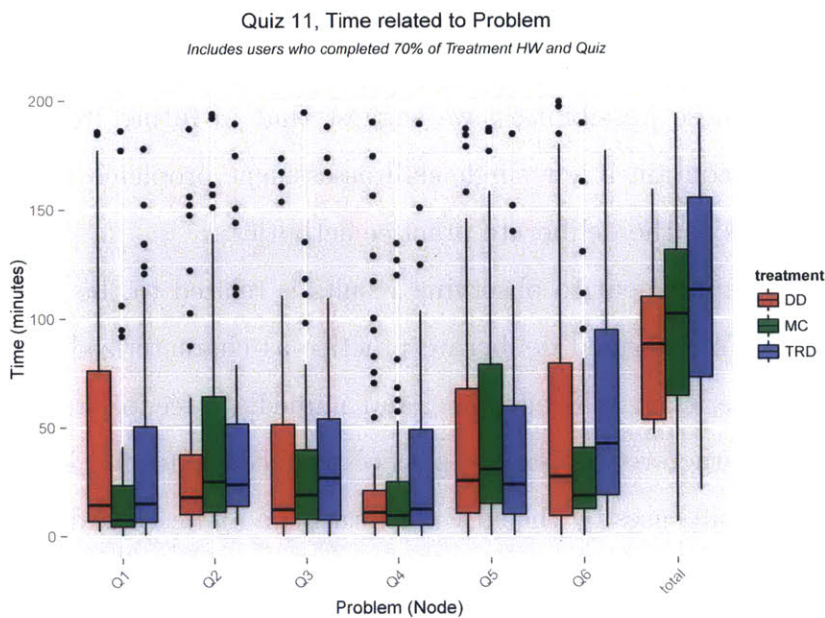


Figure 3-5: Molecular time per problem (node) on Quiz 11. Boxes show the 2nd and 3rd quartile, solid line shows median.

deliberate practice relative to traditional instruction can be made. What factors could be contributing to the lack of observed effects?

First, we note that users in 8.MReVx2014 can complete the treatment homework and quiz in any order they desire. We have included all students who completed at least 70 % of the treatment homework and quiz, irrespective of the order of completion. Second, we have not attempted to account for initial skill differences in students.

To address both of these issues, an analysis of this experiment is currently in progress that (1) postselects to include only users who completed a significant portion of the treatment before the quiz and (2) estimates user skills using item response theory (IRT) [23] rather than raw score. By using IRT, it will be possible to compare the estimated skills for each user before and after each treatment (with pre-skill based on prior student work within the course and post-skill based on common assessment). This analysis may also give insight into whether deliberate practice activities are more effective for low skill students rather than high skill students.

Third, we note that the quizzes are composed of traditional problems. It is possible

that the deliberate practice activities did not effectively train the skills we targeted, or that these skills were not the main stumbling blocks to users attempting the quiz. To investigate both these possibilities, we suggest that in future iterations of this experiment, the quiz contain a few single-skill assessment problems that target the same elementary skills as the deliberate practice activities.

A final possible impediment to observing results is related to the active population in our course. We described deliberate practice as characterized by a focus on elementary skills, repetition, self-reflection, and immediate feedback. To a certain extent, all of these characteristics (except possibly immediate feedback, which in our MOOC is provided in all cases by the edX platform) can all be viewed as personality traits rather than activity traits. In discussing deliberate practice, Chi wrote [24]

It is very difficult to say whether deliberate practice is the result of some personality or individual attributes, such as motivation or persistence, or whether it is the nature of the designed deliberate practice task that is critical for achieving elite status.

Our attempt to design deliberate practice activities can be viewed as an attempt at designing activities that arouse these traits in students. However, MOOCs have notoriously high drop-out rates and it seems entirely possible that by the tenth week of 8.MReVx2014 when this study took place, all but the most serious and motivated students had already dropped-out. In other words, students who may have benefitted from activities designed to promote deliberate practice may have already left the course.

## **3.4 Comparison of Drag and Drop vs Multiple Choice Problem Formats**

### **3.4.1 Results**

We now describe the performance of the two deliberate practice groups on the deliberate practice activities themselves. The activities for each group were designed to



contain equivalent content, and differed only in format: one group received drag-and-drop (DD) activities, the other received multiple-choice (MC) activities. Two deliberate practice activities from Unit 11 (Finding\_Error\_11.3 and Finding\_Error\_11.6) contained (unintentional) errors and are eliminated from this and subsequent analysis. Users were allowed multiple attempts on each activity. Figure 3-6 compares the first-attempt-correct and eventually-correct rates for both groups on each activity. Student's  $t$ -test was used to detect differences in the first-attempt-correct rates between groups for each activity, and the associated  $p$ -values are shown in Figure 3-6. In nine of the deliberate practice activities there was a difference in the DD and MC group first-attempt-correct rates significant at the  $p < 0.05$  level. In none of the three homework assignments was there an overall first-attempt-correct or eventually-correct difference significant at the  $p < 0.05$  level.

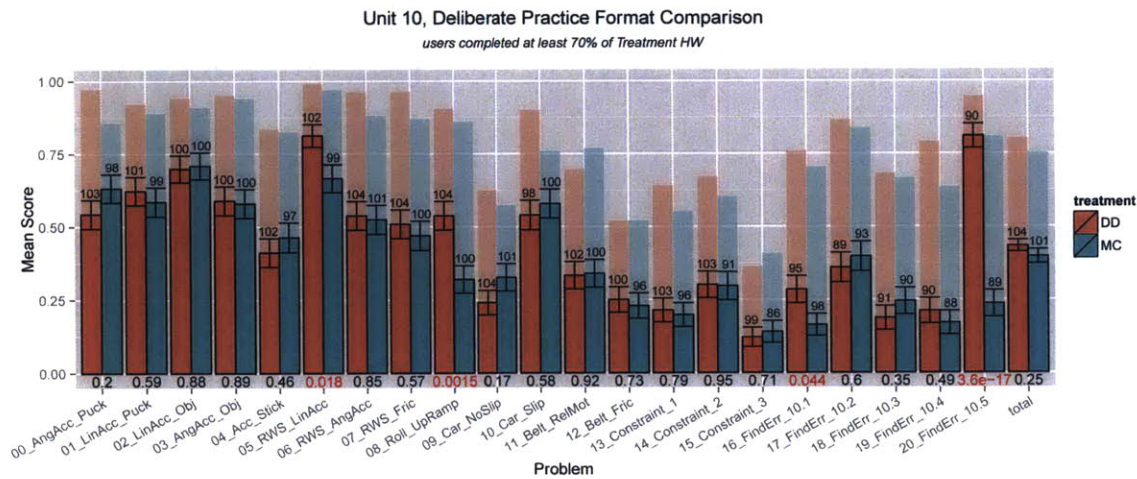


Figure 3-6: Success rates for the drag-and-drop format group (DD) and multiple-choice format group (MC) on analogous deliberate practice activities. Solid bar height indicates the fraction of students in each group who answered a particular DPA correctly on the first attempt. Error bars show one standard error in the mean. Transparent bar height indicates the fraction fraction of students who *eventually* answered each question correctly. The number above each bar indicates the number of users from each group who attempted that deliberate practice activity question. For each activity, a  $t$ -test was used to detect differences in the mean value of first-attempt-correct rates. The associated  $p$  values are shown underneath each pair of bars, with  $p$  values less than 0.05 highlighted in red.

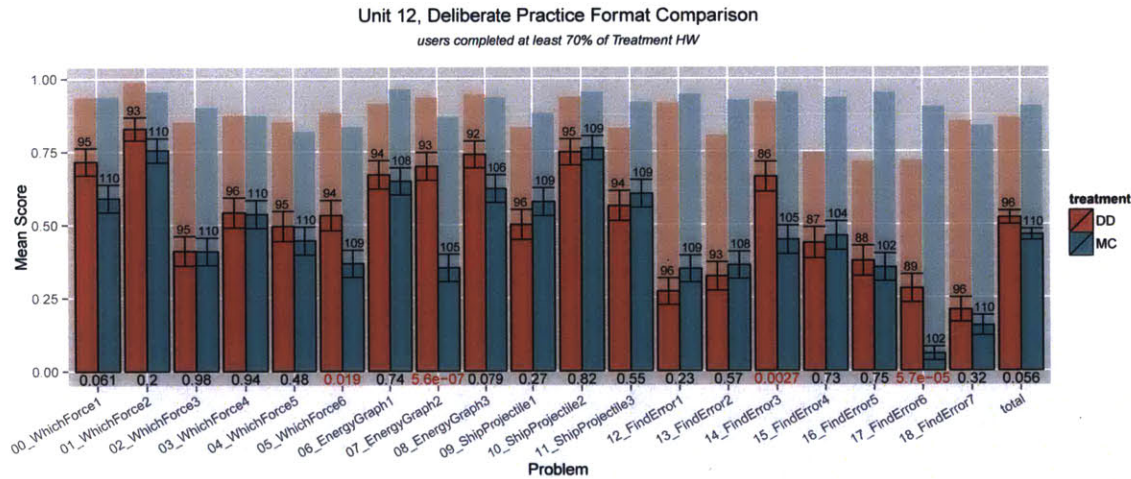
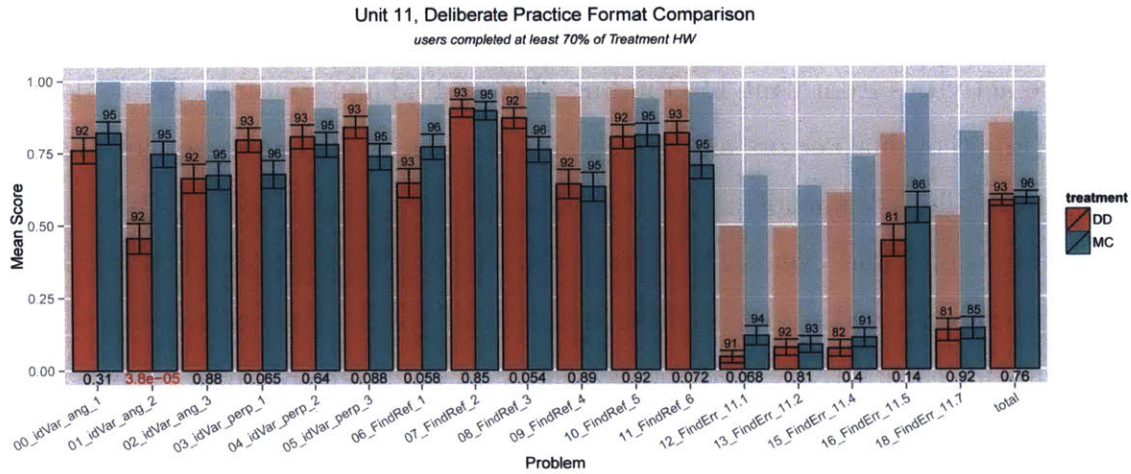


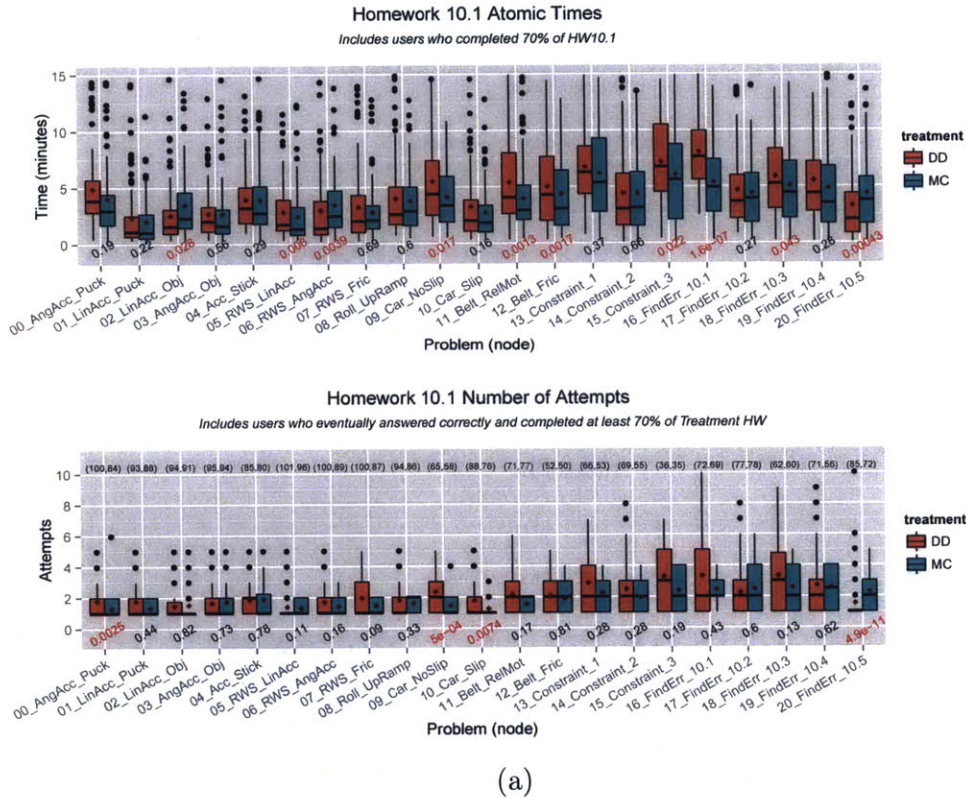
Figure 3-6: *continued*

For those users who eventually answered correctly, Figure 3-7 shows the distribution of time-until-correct and attempts-until-correct by group for each deliberate practice activity. Because each vertical within the deliberate practice activity homework assignment contained several problem nodes, we chose to measure time-per-attempt using the atomic timing algorithm, which includes only time during which the user was viewing and interacting with the activity and does not include time spent on related resources. See Appendix A for details.

We again used the Mann-Whitney  $U$ -test to detect differences in the time and



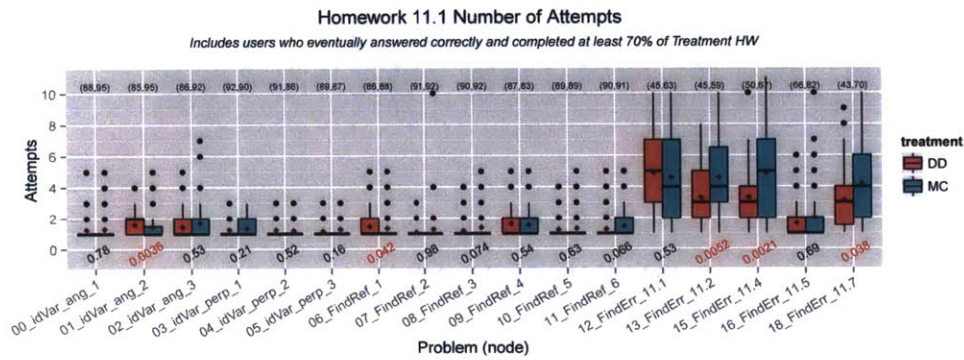
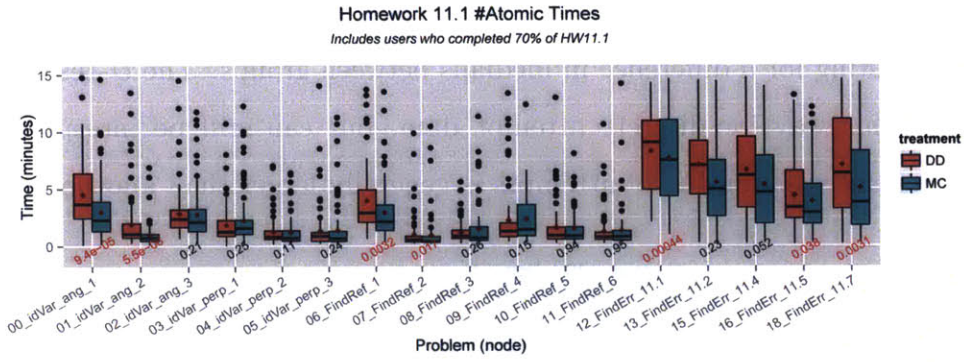
attempt distributions for each group. For continuous time data, the  $U$ -test can be interpreted as testing for a difference in medians between groups. The  $U$ -test cannot be interpreted this way for discrete attempt data, but can still be interpreted as testing whether one distribution of attempts is generally larger than another<sup>2</sup>.



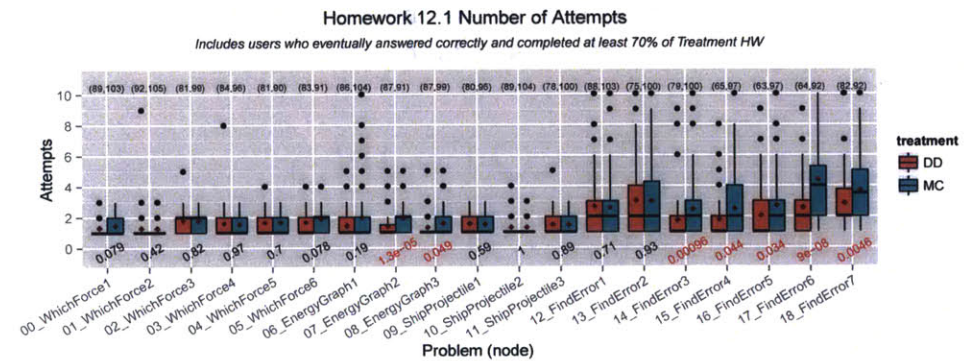
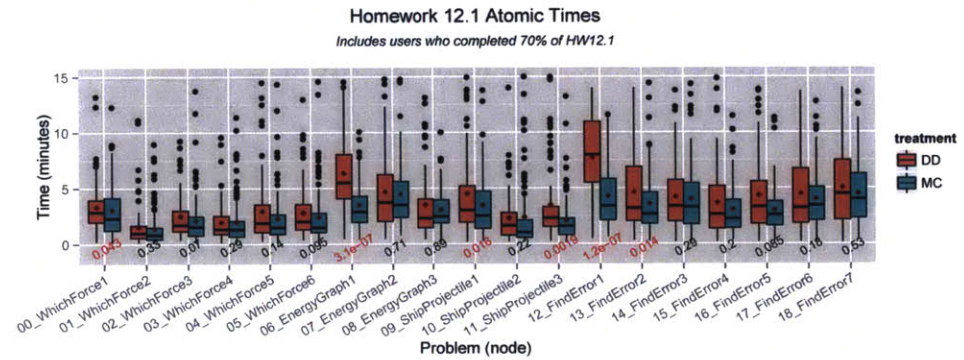
(a)

Figure 3-7: Distributions of time-until-correct and attempts-until-correct for each group on the deliberate practice activities. Times were calculated using the atomic timing algorithm. In each plot, solid lines within the box indicate distribution median, boxes represent the 2nd and 3rd quartile, and overlaid diamonds represent distribution mean. Mann-Whitney  $U$ -test  $p$  values are shown for each activity.

<sup>2</sup>In particular, for discrete data, the  $U$ -test is a rule to decide whether two random variables  $X$  and  $Y$  differ in *stochastic ordering*.  $X$  is said to be *stochastically less* than  $Y$ ,  $X \prec Y$ , if  $\Pr[X \text{ greater than } c] < \Pr[Y \text{ greater than } c]$  for all  $c \in \mathbb{R}$  [25].



(b)



(c)

Figure 3-7: *continued*



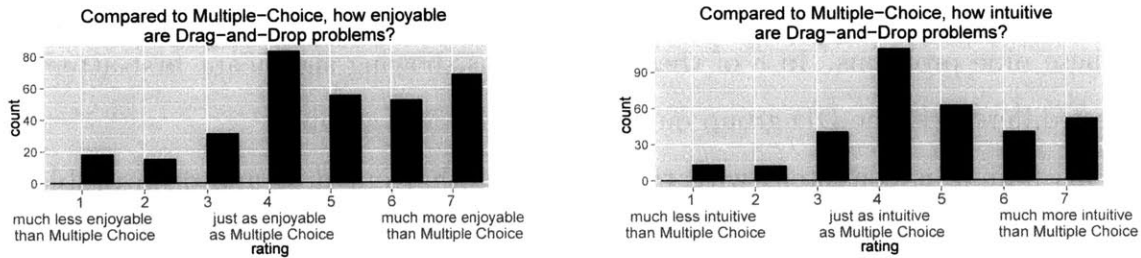


Figure 3-8: Results of survey questions comparing the drag-and-drop and multiple-choice formats.

At the end of the course (one week after Unit 12 was due), all users were given the option of participating in the course Exist Survey. Two questions on the survey asked to compare the intuitiveness and enjoyableness of the drag-and-drop format relative to multiple-choice format on a seven-point Likert scale. Before answering these questions, users were reminded that they experienced drag-and-drop format problems in one of Units 10, 11, 12 and multiple-choice format problems in another of these units. Figure 3-8 shows the results of this survey for all users who responded.

- *Enjoyableness*: Of the 322 users who responded, 26% indicated that the DD and MC formats are equally enjoyable; 54% indicated that the DD format is more enjoyable than the MC format; and 20% indicated the DD format is less enjoyable.
- *Intuitiveness*: Of the 327 users who responded, 33% indicated that the DD and MC formats are equally intuitive; 47% indicated that the DD format is more enjoyable than the MC format; and 20% indicated the DD format is less intuitive.

### 3.4.2 Discussion

Several interesting differences are visible between the drag-and-drop format and multiple-choice format deliberate practice problems.

First, we notice that although there was no significant difference in *overall* first-attempt-correct or eventually-correct rates, nine *individual* problems did exhibit first-

attempt-correct differences significant at the  $p < 0.05$  level. Table 3.6 displays data for these nine problems. In 8 of these 9 problems showing significant first-attempt-correct differences, the DD group outperformed the MC group.

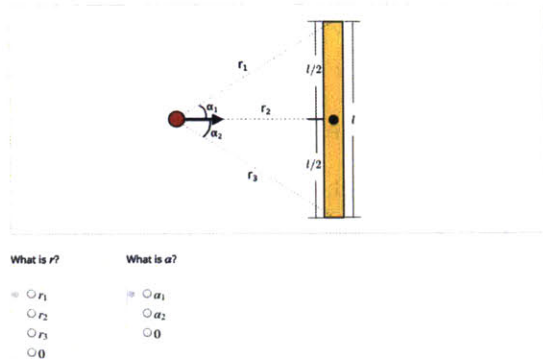
This finding is consistent with the hypothesis that our drag-and-drop problems have lower extraneous cognitive load, freeing the user’s working memory to focus on problem solving. However, as discussed in Section 3.3.1 we did not observe a significant difference between the DD and MC groups’ performance on the common assessment. Thus, although we have seen some evidence that reducing extraneous cognitive load can increase local problem solving ability, we did not observe increased learning as measured by transfer.

Table 3.6: Nine deliberate practice activities with significant first-attempt-correct

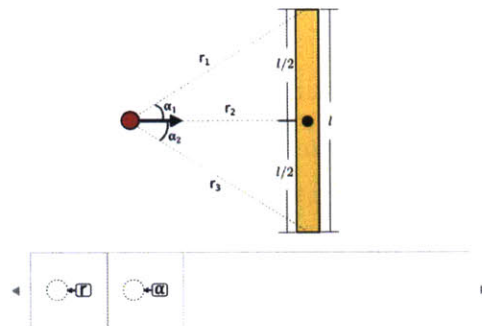
Unit	problem	DD		MC		$p$
		mean±se	$N$	mean±se	$N$	
Unit 10	05_RWS_LinAcc	0.81 ± 0.039	100	0.67 ± 0.048	99	0.018
Unit 10	08_Roll_UpRamp	0.54 ± 0.049	100	0.32 ± 0.047	100	0.0015
Unit 10	16_FindErr_10.1	0.28 ± 0.047	95	0.16 ± 0.038	98	0.044
Unit 10	20_FindErr_10.5	0.81 ± 0.041	90	0.24 ± 0.045	89	3.6e-17
Unit 11	01_idVar_ang_2	0.46 ± 0.052	92	0.75 ± 0.045	95	3.8e-05
Unit 12	05_WhichForce6	0.53 ± 0.052	94	0.37 ± 0.046	110	0.019
Unit 12	07_EnergyGraph2	0.7 ± 0.048	93	0.35 ± 0.047	100	5.6e-07
Unit 12	14_FindError3	0.66 ± 0.051	86	0.45 ± 0.049	100	0.0027
Unit 12	17_FindError6	0.28 ± 0.048	89	0.059 ± 0.023	100	5.7e-05

\*Problems for which the DD group outperformed MC are shaded gray

Of the 9 problems showing significantly different first-attempt-correct rates, “01\_id-Var\_ang\_2” stands out as the only problem in which the multiple choice group significantly outperformed the drag-and-drop group. This problem, depicted in Figure ?? was an Identify Variables activity. In the drag-and-drop version of this problem, the correct answer was to drag the  $r$  indicator to  $r_2$  and to leave the  $\alpha$  indicator in its tray (since the appropriate value of  $\alpha$ , namely  $\alpha = 0$ , is not displayed in the target figure). This was the only Identify Variables activity in which the correct answer was to leave one of the indicators unused, and we hypothesize that it was this aspect



(a) Student Task: select appropriate values of  $r$  and  $\alpha$  to be used in  $L = rmv \sin \alpha$  by marking multiple-choice radio buttons. Correct answer is  $(r, \alpha) = (r_2, 0)$ .



(b) Student Task: select appropriate values of  $r$  and  $\alpha$  to be used in  $L = rmv \sin \alpha$  by dragging indicator to figure. Correct answer is to indicate  $r_2$  and leave the  $\alpha$  indicator in its tray.

Figure 3-9: Comparison of multiple-choice and drag-and-drop format for a variable identification deliberate practice activity.

of the activity that made the problem artificially hard for the DD group. In future versions of this activity, we will implement a more active mechanism by which the user indicates the idea “ $\alpha = 0$ ”, e.g., by having two more draggable objects,  $\boxed{\alpha = 0}$  and  $\boxed{r = 0}$ .

Although the DD group performed significantly better than the MC group on several problems *on the first attempt*, the MC group had higher *eventually-correct* rates in many cases, especially in Units 11 and 12 and especially on the Finding Error activities. (In all cases, the multiple choice group had equal or fewer attempts than the drag-and-drop group.) This effect probably results from two causes. First, multiple choice problems are susceptible to guessing strategies when multiple attempts are given since the number of options is finite. In contrast, many drag-and-drop problems have a very large of options, since grading is based on the coordinates of the draggable objects. Second, as discussed in Section 3.2.3, the multiple choice group receives more feedback per attempt.

The time-on-task data also exhibits trends. First, we observe that when multiple problems of a similar type are given (e.g., Identify Variables, or Find the Reference Point activities, or multiple Which Force activities) in a row, the time-per-problem

tends to decrease. This makes sense, as users are acquiring a familiarity with the problem type.

More interestingly, we observe that on several problems, the drag-and-drop group spends *more time per problem* and uses *fewer attempts* to get the problem correct. These trends are especially apparent in the Finding Error activities used in Units Eleven and Twelve. For the 12 finding error activities considered in these units:

- The DD median time was larger in 11 of the 12 cases, and the effect was significant at  $p < 0.05$  in 5 of these cases ( $U$ -test).
- The DD median number of attempts was smaller in 8 of the 12 cases, and the attempt distributions were significantly different at  $p < 0.05$  in all 8 of these cases ( $U$ -test).
- Additional analysis shows the DD group median time per attempt was larger in all 12 cases, and significant at  $p < 0.05$  ( $U$ -test) in 10 of the 12 cases.

Time-on-task is often used as a measure of cognitive load, and we may suspect that the DD group actually experienced higher cognitive load during these Finding Error activities. However, since the DD group used fewer attempts in many cases, it seems that this increased cognitive load was germane to problem solving rather than extraneous cognitive load related to problem format. One explanation for these findings is that the drag-and-drop group found their Finding Error activities to be more interesting and more interactive, and hence were willing to engage more with the material. This interpretation seems to be supported by the exit survey, on which a majority of respondents indicated they found drag-and-drop problems to be more enjoyable and more intuitive.

### 3.5 Conclusions

We observed some evidence that drag-and-drop activities have higher success rates than their multiple-choice counterparts, suggesting that drag-and-drop can be used

to reduce extraneous cognitive load. Users tended to spend more time on drag-and-drop problems, especially of the Finding Error type, but often arrived at the correct answer in fewer attempts. Survey results indicate that a majority of users find drag-and-drop problems more intuitive and more enjoyable, possibly explaining why users are willing to spend more time per attempt on activities.

We observed no significant evidence that deliberate practice activities, rendered in the drag-and-drop or multiple-choice format, result in better learning as measured by transfer on a post-assessment. We suggest that future work into the role of deliberate practice on learning in MOOCs should focus on low-skill populations in the hopes of observing a stronger effect. Also, we suggest that large-scale learning experiments be carried out within the early few weeks of a MOOC when the low-skill population is presumably larger. Moreover, learning experiments early in a MOOC seem to have a better chance of observing retention effects.



# Chapter 4

## Experiment: Does Pre-Test Feedback enhance Post-Test performance?

### 4.1 Background

A standard method for measuring learning is to administer the same assessment before (pre-test) and after (post-test) instruction. This technique has been used extensively in education research in residential environments and was recently used in a MOOC environment to demonstrate equal learning among different student cohorts in 8.MReVx2013 [3].

The pre-/post-testing in 8.MReVx2013 differed from traditional residential pre-/post-testing in several significant ways:

1. users were given multiple attempts on each problem and were given correct/incorrect feedback after each attempt;
2. users were able to see the correct answer to a problem after finishing the problem;
3. the pre- and post-tests were “open”, i.e., users could take advantage of outside resources (for the post-test, this includes instruction material that had been released within 8.MReVx).

4. Except for a due date two-weeks after release, no time-limit was imposed on the pre-test or the post-test;

How important are these differences between the MOOC and residential pre-/post-test methodologies? 8.MReVx2014 included an experiment to investigate whether feedback and the ability to see the correct answers on the pretest would enhance student performance on identical problems on the post-test twelve weeks later.

## 4.2 Study Setup

Figure 4-1 shows the setup for the pre/post-test memory experiment. At the beginning of the course, all students had the opportunity to complete a pre-test and at the end of the core all student had the opportunity to complete a post-test. All students received the same post-test; twos different versions of the pre-test were given. The post-test contained fifteen separate problems corresponding to 23 edX input fields (some problems have multiple parts). Post-test problems can be grouped into four item categories according to which group saw the items on the pre-test:

- Two problems (3 edX input fields, i1–i3) appeared only on pre-test version A
- Two problems (4 edX input fields, i4–i7) appeared only on pre-test version B
- Nine problems (12 edX input fields, i8–i19) appeared on both pre-test versions A and B
- Two problems (4 edX input fields, i20–i23) were unique to the post-test, and did not appear on either pre-test version.

The problems appeared in different orders on the pre-test and post-test. The numbering of input fields in Figure 4-1 does not correspond to the order in which they appeared on either test.

Users were allowed multiple attempts (usually 2–4) on the pre/post-test problem nodes and were not penalized for using multiple attempts. The pre-test was ungraded; the post-test was displayed as the course “final exam” and was worth 9% of each user’s



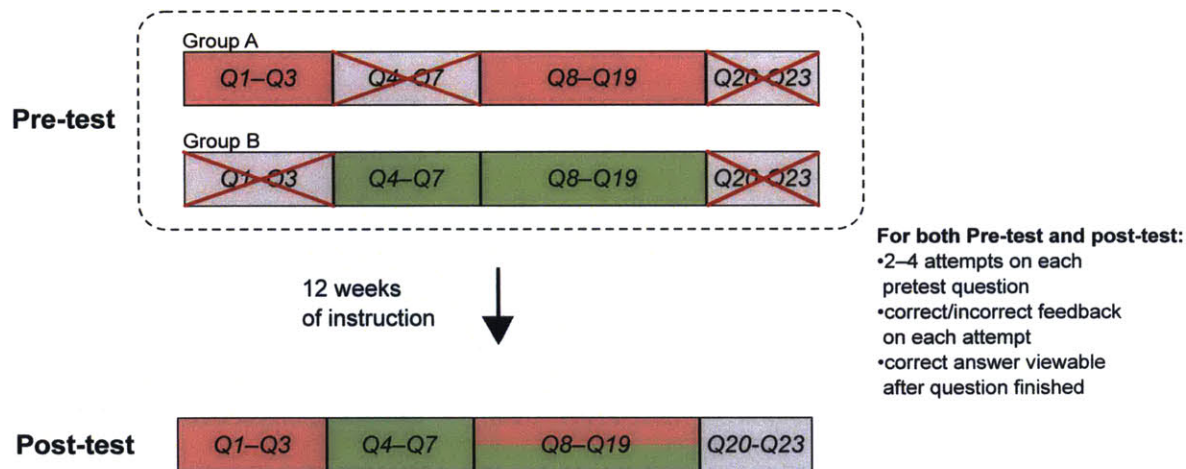


Figure 4-1: Setup for the pre/post-test memory experiment. All students see the same common post-test after 12 weeks of instruction, but two slightly different versions of the pre-test were given at the beginning of the course.

total grade. Users received correct/incorrect feedback on each input field after hitting the `check` button for each attempt, and were able to view the correct answer after either using all their attempts or correctly answering the question.

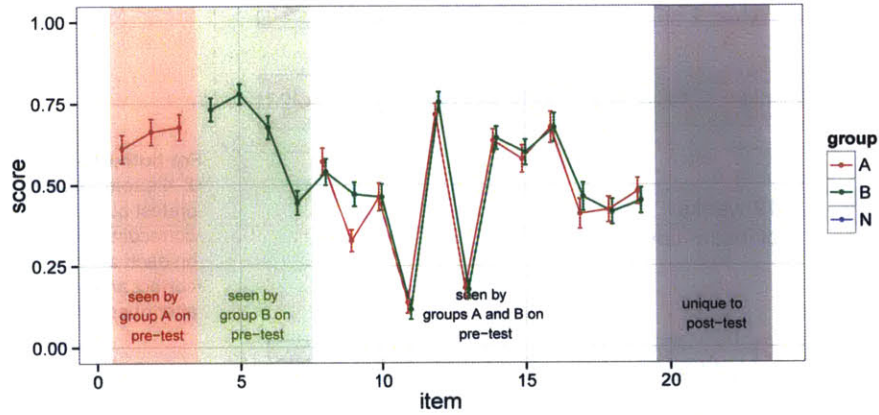
Twelve weeks of instruction were available between the pre-test and the post-test. The pre-test was hidden from students after the second week of instruction, and was not visible to students after that time.

### 4.3 Results and Discussion

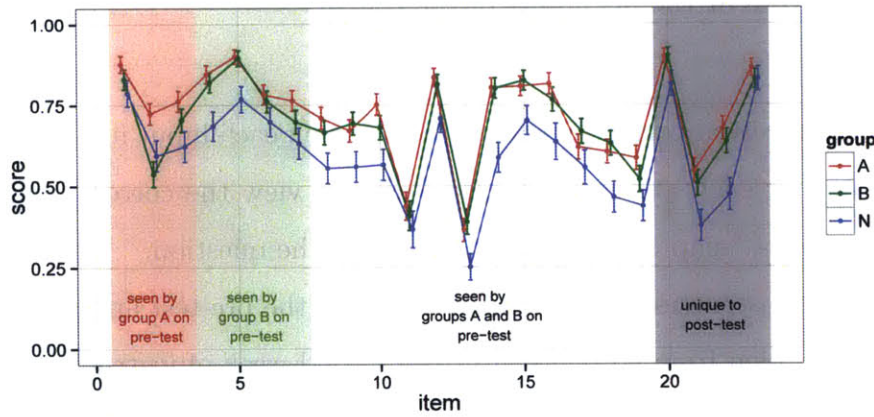
A total of 516 users attempted the 8.MReVx2014 post-test; not all users attempted all problems on the post-test. On average, users attempted 85% of problems on the post-test. Each user was assigned to either version A or version B of the pre-test, but not all of these 516 users completed the pre-test.

To avoid score saturation, we use first-attempt-correct rate to quantify performance on the pre- and post-tests. Figure 4-2 shows the mean first-attempt-correct rates on each pre- and post-test input field for three groups of students:

- **Group A** ( $N = 198$ ) includes of users who were assigned to version A of the pre-test and attempted *at least some* items on the pre-test



(a) pre-test scores by group



(b) post-test scores by group

Figure 4-2: Pre-test scores for groups A and B, and post-test scores for groups A, B, and N. Error bars show one standard error in the mean.

Table 4.1: First-Attempt-Correct Rates by user group and item category for post-test

item category	Group A				Group B				Group N			
	<i>N</i>	mean	sd	sem	<i>N</i>	mean	sd	sem	<i>N</i>	mean	sd	sem
pre-test A	165	0.788	0.268	0.021	146	0.705	0.275	0.023	95	0.684	0.327	0.034
pre-test B	165	0.830	0.239	0.019	151	0.816	0.229	0.019	92	0.715	0.307	0.032
pre-test A, B	112	0.714	0.172	0.016	106	0.706	0.191	0.019	56	0.616	0.235	0.031
unique to post-test	169	0.750	0.245	0.019	146	0.723	0.249	0.021	90	0.656	0.279	0.029

- **Group B:** ( $N = 186$ ) includes of users who were assigned to version B of the pre-test and attempted *at least some* items on the pre-test
- **Group N:** ( $N = 132$ ) includes of users who were assigned to version A or B of the pre-test but did not attempt *any* items on the pre-test.

The performance of each user group (A, B, N) on each of these item category (seen by A on pre-test, seen by B on pre-test, seen by A and B on pre-test, seen by neither on pretest) is summarized in Table 4.1. The mean first-attempt-correct rate for each user group on each item category is calculated on the subset of users that attempted all items in that particular item category.

The most striking feature of Figure 4-2a is that users who attempted the pre-test (either group A or B) consistently had higher first-attempt-correct rates on the post-test than users who did not attempt the pre-test (group N). This is true across *all* item categories, including items that were unique to the post-test. This suggests that the stronger performance of groups A and B is not due to a memory effect whereby students learned from the pre-test, but rather is due to a self-selection effect where weaker students in group N chose not to participate in the pre-test.

Because users were assigned randomly to pretest version A or pretest version B, we expected that groups A and B would perform similarly on post-test items seen common to both pre-test versions or unique to the post test. We expected that user group A would have a slight advantage on the set of items unique to pre-test version A and that group B would have a slight advantage on the set of items unique to pre-test version B.

In contrast to our expectation, group A had higher mean first-attempt-correct rates on *all* item categories. This suggests that students in group A are somewhat more skilled than students in group B. To determine whether group A had an advantage on the items that appeared on pre-test version A, we compare the difference in mean first-attempt correct rates between groups A and B on all item categories, shown in Table 4.2.

Table 4.2 shows that group A did somewhat better across all item categories on

Table 4.2: Difference in post-test first-attempt-correct rates for Groups A and B

item category	post-test first-attempt-correct		
	mean <sub>A</sub> – mean <sub>B</sub>	sem <sub>A</sub>	sem <sub>B</sub>
pre-test A	0.082	0.021	0.023
pre-test B	0.014	0.019	0.019
pre-test A, B	0.0075	0.016	0.019
unique to post-test	0.027	0.019	0.021

the post-test and that the difference was larger on items they had seen on the pretest. We take the difference in differences

$$\left( \text{mean}_{\text{pre-test A}}^{\text{group A}} - \text{mean}_{\text{pre-test A}}^{\text{group B}} \right) - \left( \text{mean}_{\text{pre-test A,B}}^{\text{group A}} - \text{mean}_{\text{pre-test A,B}}^{\text{group B}} \right) = 0.075 \quad (4.1)$$

as a measure of A’s advantage due to exposure to problems on the pre-test, rather than do A’s apparent higher overall skill. Although this difference in differences is positive, it is significant only at  $p = 0.058$  (see Appendix B). Thus it seem that, averaged over all items, transfer from pre-test to post-test did not contribute significantly to group A’s higher score.

However, one item on the post-test, i2, showed an especially large difference in performance between groups A and B, and is worthy of particular discussion. This item was seen only by group A on the pre-test and group A performed over three standard deviations better than group B on this item (Table 4.3). Thus transfer from pre-test to post-test may have contributed to group A’s superior performance *on this particular item*, and we suggest that the item’s peculiar format may explain why transfer was observed for this item alone. Item i2 was a “problem-decomposition” activity in which the user is presented with a complex physics problem and asked whether the problem is best broken into 1, 2, or 3 subparts. This type of activity is rare in 8.MReVx2014: excluding the pre- and post-tests, only two such problem-decomposition activities appeared in the course<sup>1</sup>. Thus, group A would have been

<sup>1</sup>The homework assignment for Unit 8: Mechanical Energy and Work contained two problem-decomposition activities.

exposed to 50 % more items of this format than group B upon entering the post-test. In contrast, for items of more standard format (e.g., numerical calculation, symbolic, or multiple choice conceptual) both groups would have been exposed to roughly equal amounts of such problems throughout the course.

Table 4.3: Post-test performance on Item i2

Group A			Group B			Difference	
<i>N</i>	mean	sem	<i>N</i>	mean	sem	mean	sem
181	0.724	0.0333	160	0.538	0.0395	0.186	0.0517

## 4.4 Conclusions

Our experiment showed little evidence for enhancement of post-test scores due to students seeing the same items on the pre-test, even though the pre-test gave correct/incorrect feedback and provided students with the correct answer after finishing the item. This bodes well for those who plan to use pre-/post-testing to measure learning in MOOCs where experiment designers may not have complete control over the platform.



# Appendix A

## Brief Description of the Atomic and Molecular Timing Algorithms

To estimate the amount of time students spend working on each problem in the 8.MReVx2014 our group analyzes the edX JSON log. Here we give a brief description of the “atomic” and “molecular” timing algorithms written by Giora Alexandron. The atomic timing algorithm estimates the amount of time students spend working on a particular problem without the use of other in-course resources, while the molecular timing algorithms estimates the amount of time a student spends working on a problem *plus* related in-course resources.

### A.1 Time interacting with problem (“atomic” time)

The edX log files contain records of several types of events including page and problem load events, problem check events, and show solution events. One challenge to measuring the time spent on a particular problem without including related in-course resources is that edX log files do not record when a user leaves a particular page. A large approximation used by the atomic timing algorithms is that the user only ever has a single edX page open at a time. (Users with multiple browser tabs or browser windows open on 8.MReVx2014 clearly violate this approximation.) Under this approximation, whenever the user enters a new page, the user has stopped viewing the

previous page; both timing algorithms begin by adding artificial [LEAVE PAGE] events to the log files.

The “atomic” timing algorithm first breaks the log file into disjoint time intervals.

- Start event: Every interval begins at at [problem get] or [problem check] event.
- End event: Once an interval has started, the interval ends at the next [LEAVE PAGE] or [problem\_check] event.

It is assumed that during this interval, the user is working on *some* problem on the vertical containing the problem that triggered the start event. The time interval is then attributed to the next problem that has a [problem\_check] event. This is illustrated in Figure A-1

Additionally, very short and very long intervals are discarded. If an interval has length less than 10 seconds, we assume the user is navigating edX rather than working on a problem. If an interval is greater than 30 minutes in length, it is assumed the user is not actively engaged with the content. Such intervals are assigned a duration of 0 s instead of their actual length.

	<b>timestamp</b>	<b>event type</b>	<b>resource</b>
	⋮		
<i>We infer that the user is working on a problem, but has decided not to check it. The elapsed time is attributed to the next check event. (problem_a)</i>	t1	[goto page]	[vertical_1_with_problems]
	t2	[problem_get]	[problem_a]
	t3	[problem_get]	[vertical_1_problem_part_b]
	t4	[LEAVE PAGE]	[vertical_1_with_problem]
<i>this interval is ignored by the atomic timing algorithm</i>	t5	[goto page]	[vertical_2_with_html]
	t6	[LEAVE PAGE]	[vertical_2_with_html]
	t8	[goto_page]	[vertical_1_with_problem]
<i>time attributed to problem_a</i>	t9	[problem_get]	[problem_a]
	t10	[problem_get]	[problem_b]
<i>time attributed to problem_a</i>	t11	[problem_check]	[problem_a]
<i>time attributed to problem_b</i>	t12	[problem_check]	[problem_a]
	t13	[problem_check]	[problem_a]
	⋮		

Figure A-1: Annotated schematic of a log file analyzed by the atomic timing algorithm.



## A.2 Total time on problem and related resources (“molecular” time)

The “molecular” timing algorithm attempts to estimate the total time spent on a particular problem and related resources by taking the time difference between the last [problem\_check] and first [problem\_get] events for a particular problem. Any period of user activity between these two events that lasts less than 10 seconds is ignored (assumed to correspond to site navigation) and any period of user activity longer than 30 minutes is also ignored.

Because the [problem\_check] events for all problems on a given vertical occur simultaneously, the molecular timing algorithm does not work well if multiple unrelated problems occur on a single vertical. For this reason, most verticals within 8.MReVx2014 contain only one edX problem node or multiple problem nodes that are sub-parts of the same physics problem, and hence time spent on part is reasonably attributed as time “related to” another part.



# Appendix B

## Calculating Advantage Due to Exposure on Pre-test

To quantify the advantage gained by group A on the post-test problems that appeared on version A of the pre-test, we compare the difference in first-attempt-correct rates between groups A and B on the item categories “seen by A (on pre-test)” and “seen by A and B (on pre-test)”. Let

- $X_{A \text{ on } A}$  denote the first-attempt-correct rate averaged over users in group A and over the post-test items that were seen by group A and not by B on the pre-test.
- $X_{B \text{ on } A}$  denote the first-attempt-correct rate averaged over users in group B and over the post-test items that were seen by group A and not by B on the pre-test.
- $X_{A \text{ on } Com}$  denote the first-attempt-correct rate averaged over users in group A and over the post-test items that were seen by both groups on the pre-test.
- $X_{B \text{ on } Com}$  denote the first-attempt-correct rate averaged over users in group B and over the post-test items that were seen by both groups on the pre-test.

We assume that each of these random variables is independent from one another and that they are normally distributed due to the averages being taken over relatively

large sample sizes (at least 100 users in each case; see Table 4.2.

We take the difference  $X_{A \text{ on } A} - X_{B \text{ on } A}$  as measuring two effects:

1. The advantage of group A on items A due to the difference in skill between groups A and B
2. The advantage of group A on items A due to exposure to these items on the pre-test

and take  $X_{A \text{ on } Com} - X_{B \text{ on } Com}$  as a measure of group A's advantage due to difference in skill alone. To isolate the effect of exposure to items on the pre-test, we consider the difference

$$X_{A's \text{ pre-test advantage}} = (X_{A \text{ on } A} - X_{B \text{ on } A}) - (X_{A \text{ on } Com} - X_{B \text{ on } Com}). \quad (\text{B.1})$$

which is also normally distributed. We estimate the mean and variance of  $X_{A's \text{ pre-test advantage}}$  using the data in Table B.1:

$$\overline{X}_{A's \text{ pre-test advantage}} = (\overline{X}_{A \text{ on } A} - \overline{X}_{B \text{ on } A}) - (\overline{X}_{A \text{ on } Com} - \overline{X}_{B \text{ on } Com}) = 0.075 \quad (\text{B.2})$$

$$s_{A's \text{ pre-test advantage}} = \sqrt{s_{A \text{ on } A}^2 + s_{B \text{ on } A}^2 + s_{A \text{ on } Com}^2 + s_{B \text{ on } Com}^2} = 0.040. \quad (\text{B.3})$$

We wish to test the hypothesis that exposure to items on the pre-test affects post-test performance on the same items. To that end, we construct the the  $t$ -statistic [22]

$$t = \frac{\overline{X}_{A's \text{ pre-test advantage}}}{s_{A's \text{ pre-test advantage}}} = 1.89. \quad (\text{B.4})$$

Under the null hypothesis ( $\mu_{A's \text{ pre-test advantage}} = 0$ ), a  $t$ -statistic at at least this large in magnitude occurs with probability

$$p = \int_{|u|>1.89} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = 0.058, \quad (\text{B.5})$$

where we used the standard normal distribution to approximation the  $t$  distribution due to our fairly large sample size.

Table B.1: First-Attempt-Correct Rates by user group and item category for post-test

item category	Group A				Group B			
	<i>N</i>	mean	sd	sem	<i>N</i>	mean	sd	sem
pre-test A	165	0.788	0.268	0.021	146	0.705	0.275	0.023
pre-test B	165	0.830	0.239	0.019	151	0.816	0.229	0.019
pre-test A, B	112	0.714	0.172	0.016	106	0.706	0.191	0.019



# Bibliography

- [1] Lori Breslow and DE Pritchard. Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice . . .*, (March 2012):13–25, 2013.
- [2] D Seaton, Isaac Chuang, P Mitros, and D Pritchard. Who Does What in a Massive Open Online Course? *Communications of the ACM*, 57(4):58–65, 2014.
- [3] Kimberly F Colvin, John Champaign, Alwina Liu, Qian Zhou, Colin Fredericks, and David E Pritchard. Learning in an introductory physics MOOC: All cohorts learn equally, including an on-campus class, August 2014.
- [4] KA Ericsson, RT Krampe, and C Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3):363–406, 1993.
- [5] John Sweller, Paul Ayres, and Slava Kalyuga. *Cognitive Load Theory*. 2011.
- [6] WG Chase and HA Simon. Perception in chess. *Cognitive psychology*, 61:55–81, 1973.
- [7] D E Egan and B J Schwartz. Chunking in recall of symbolic drawings. *Memory & cognition*, 7(2):149–58, March 1979.
- [8] Michelene T H Chi, R Glaser, and E Rees. Expertise in Problem Solving. In RJ Sternberg, editor, *Advances in the psychology of human intelligence*. 1981.

- [9] Michelene T H Chi, PJ Feltovich, and Robert Glaser. Categorization and representation of physics problems by experts and novices\*. *Cognitive science*, 5(2):121–152, 1981.
- [10] R T Krampe and KA Ericsson. Maintaining excellence: deliberate practice and elite performance in young and older pianists. *Journal of experimental psychology. General*, 125(4):331–59, December 1996.
- [11] Neil Charness, Michael Tuffiash, Ralf Krampe, Eyal Reingold, and Ekaterina Vasyukova. The role of deliberate practice in chess expertise. *Applied Cognitive Psychology*, 19(2):151–165, March 2005.
- [12] Anique B H de Bruin, Niels Smits, Remy M J P Rikers, and Henk G Schmidt. Deliberate practice predicts performance over time in adolescent chess players and drop-outs: a linear mixed models analysis. *British journal of psychology (London, England : 1953)*, 99(Pt 4):473–97, November 2008.
- [13] KA Ericsson. The influence of experience and deliberate practice on the development of superior expert performance. In K. Anders Ericsson, Neil Charness, Paul J. Feltovich, and Robert R. Hoffman, editors, *Cambridge handbook of expertise and expert performance*. Cambridge University Press, Cambridge, 2006.
- [14] KA Ericsson. Discovering deliberate practice activities that overcome plateaus and limits on improvement of performance. *International Symposium on Performance . . .*, 2009.
- [15] Joseph Baker and Bradley Young. International Review of Sport and Exercise Psychology 20 years later : deliberate practice and the development of expertise in sport. *International Review of Sport and Exercise Psychology*, 7(1):135–157, 2014.
- [16] Ronald T. Kellogg and Alison P. Whiteford. Training Advanced Writing Skills: The Case for Deliberate Practice. *Educational Psychologist*, 44(4):250–266, October 2009.



- [17] Kimberly a Gifford and Leslie H Fall. Doctor coach: a deliberate practice approach to teaching and learning clinical skills. *Academic medicine : journal of the Association of American Medical Colleges*, 89(2):272–276, February 2014.
- [18] Elizabeth A EA Hunt, Jordan M JM Duval-Arnould, Kristen L Nelson-McMillan, Jamie Haggerty Bradshaw, Marie Diener-West, Julianne S Perretta, and Nicole A Shilkofski. Pediatric resident resuscitation skills improve after “Rapid Cycle Deliberate Practice” training. *Resuscitation*, 85(7):945–51, July 2014.
- [19] Louis Deslauriers, Ellen Schelew, and Carl Wieman. Improved learning in a large-enrollment physics class. *Science (New York, N. Y.)*, 332(6031):862–4, May 2011.
- [20] Fred G. Paas. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84(4):429–434, 1992.
- [21] Rohani a. Tarmizi and John Sweller. Guidance during mathematical problem solving. *Journal of Educational Psychology*, 80(4):424–436, 1988.
- [22] David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*, volume 46. 2004.
- [23] R. J. de Ayala. *The Theory and Practice of Item Response Theory*. The Guilford Press, 1 edition, 2008.
- [24] Michelene T H Chi. Theoretical Perspectives, Methodological Approaches, and Trends in the Study of Expertise. *EXPERTISE IN MATHEMATICS INSTRUCTION: AN INTERNATIONAL PERSPECTIVE*, pages 17–39, 2011.
- [25] Michael P. Fay and Michael A. Proschan. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules, 2010.