

## MIT Open Access Articles

*Learning Experiments Using AB Testing at Scale*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Christopher Chudzicki, David E. Pritchard, and Zhongzhou Chen. 2015. Learning Experiments Using AB Testing at Scale. In Proceedings of the Second (2015) ACM Conference on Learning @ Scale (L@S '15). ACM, New York, NY, USA, 405-408.

**As Published:** <http://dx.doi.org/10.1145/2724660.2728703>

**Publisher:** Association for Computing Machinery (ACM)

**Persistent URL:** <http://hdl.handle.net/1721.1/99202>

**Version:** Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

**Terms of use:** Creative Commons Attribution-Noncommercial-Share Alike



---

# Learning Experiments Using AB Testing at Scale

**Christopher Chudzicki**

Department of Physics  
Massachusetts Institute of Technology  
77 Massachusetts Avenue.  
Cambridge, MA 54321 USA  
chudzick@mit.edu

**Zhongzhou Chen**

Department of Physics  
Massachusetts Institute of Technology  
77 Massachusetts Avenue.  
Cambridge, MA 54321 USA  
zchen22@mit.edu

**David E. Pritchard**

Department of Physics  
Massachusetts Institute of Technology  
77 Massachusetts Avenue.  
Cambridge, MA 54321 USA  
dpritch@mit.edu

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [Permissions@acm.org](mailto:Permissions@acm.org).

LAK '15, March 16 - 20, 2015, Poughkeepsie, NY, USA  
Copyright is held by the owner/author(s). Publication rights licensed to ACM.  
ACM 978-1-4503-3417-4/15/03...\$15.00  
<http://dx.doi.org/10.1145/2723576.2723582>

**Abstract**

We report the one of the first applications of treatment/control group learning experiments in MOOCs. We have compared the efficacy of deliberate practice—practicing a key procedure repetitively—with traditional practice on “whole problems”. Evaluating the learning using traditional whole problems we find that traditional practice outperforms drag and drop, which in turn outperforms multiple choice. In addition, we measured the amount of learning that occurs during a pretest administered in a MOOC environment that transfers to the same question if placed on the posttest. We place a limit on the amount of such transfer, which suggests that this type of learning effect is very weak compared to the learning observed throughout the entire course.

**Introduction**

The most trustworthy inferences in a complex domain like medicine or education are drawn from experiments comparing outcomes in treatment and control groups that are randomly selected from the same population. Reliability is then largely dependent on statistical uncertainty implying that large samples can support conclusions with greater certainty and/or on smaller effects. Thus treatment/control experiments (A/B experiments) in MOOCs (Massive Open Online Courses) can leverage the large number of students to study small effects such as learning from particular

### **The 8.MReVx MOOC**

is designed for those with existing knowledge of Newtonian mechanics. It concentrates on problem-solving and uses a pedagogy called Modeling Applied to Problem Solving [5].

The 2014 iteration of 8.MReVx contains 12 weeks of material, generally a single unit that contains instruction material, homework, and weekly quiz. There is a pre-test given on the first week of the course, and a post-test given at the end of the course. Most graded problems in the course allow multiple attempts, open response homework problems typically allowed 10 attempts, except only ~3 on tests. Multiple choice problems had appropriately fewer attempts also. We had ~11,000 participants in the course, with just over 500 receiving certificates.

interventions. We will discuss preliminary analysis of two of the seven A/B experiments we've done.

### **AB Experiments in MOOCs**

The edX platform has the ability to implement A/B experiments [1] in which the user population is partitioned into two or more groups and each group is given a different version of course material. Some important aspects of edX AB experiments are:

- All students who elect to take the MOOC are randomly assigned to two or more groups that receive different instructional resources for a given experiment.
- The outcomes are evaluated by giving the same questions to both groups after the instruction is completed.

We report here the preliminary results on two of our seven experiments.

### **Experiment 1: Deliberate Practice and Interactive Problem Format**

#### *Introduction:*

The goal of this study is to see if we could design new types of online homework problems that are more effective in developing problem solving abilities. Our design is based on two related ideas: deliberate practice and cognitive load theory. Work by Ericsson in the 1990s [2] showed that deliberate practice—characterized by a singular focus on elementary skills (procedures), repetition, immediate feedback, and self-reflection—was especially important for the development of expertise. In addition, cognitive load theory [3], [4] suggests that enhanced learning is achieved by reducing extraneous cognitive load, freeing the learner's working memory to focus on the most salient aspects of the activities. The edX platform

provides a chance to design new, more intuitive problem formats with rapid feedback that follow those principles much better than traditional homework.

This experiment uses three groups to answer two separate but related questions. First, we investigate whether deliberate practice activities can build physics expertise more efficiently than traditional practice involving traditional whole problems. Second, we vary the problem format of the deliberate practice activities, comparing the common multiple-choice problems with informationally equivalent "drag-and-drop" problems that minimizes the extraneous cognitive load of multiple-choice incurred by having to match each choice item with the problem body.

#### *Methods:*

Students are randomly assigned to three groups (A, B, or C) which receive one of three different treatments during each of three successive graded units (10, 11, and 12) of our MOOC (8.MReVx on the edX.org platform). The control treatment is traditional (TRD) "whole problem" homework problems; the two variations of the deliberate practice activities differ only in format (multiple choice, MC, vs drag-and-drop, DD). In order to treat all participants in our study equally, the treatment assigned to each group rotates between units. For example, Group C received the TRD problems in Unit 10, deliberate practice problems in the DD format during Unit 11, and deliberate practice problems in the MC format during Unit 12.

A common quiz is given to all three groups together with the homework. The quiz consists of traditional problems only (mostly numeric/symbolic response).

*The Traditional Problems* are a mix of conceptual and symbolic problems that mirror homework in a standard physics course.

*The Deliberate Practice* problems consisted of short problems targeting specific skills used in solving whole problems. Because each deliberate practice problem takes a shorter amount of time, we are able to include 4 times as many problems, while taking roughly the same total time for students to complete.

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

We created two versions of the deliberate practice treatment. One version (MC) uses the standard multiple-choice (MC) problem formats while The second version (DD) uses edX's drag-and-drop (DD) problem format which is much more interactive and should reduce extraneous cognitive load.

### Analysis:

In order to guarantee that the users we consider interacted with the treatment homework to a significant extent, we restrict our attention to only those who completed at least 70% of the treatment homework and at least 70% of the common quiz. Because the vast majority of users completed either very little or almost all of the activities in these units, our results are relatively insensitive to the particular cut-off used. This cut-off leaves a total of 219 students for Unit 10, 205 students for Unit 11, and 280 students for Unit 12.

Unit 10					Unit 11				Unit 12				Total		Differences			
Treatment	Group	N	mean	sd	Group	N	mean	sd	Group	N	mean	sd	mean	se	mean	sd	Z-score	p-value
DD	A	78	61%	23%	C	72	47%	22%	A	87	59%	22%	55.7%	1.5%				
MC	B	71	56%	24%	B	64	43%	24%	C	103	61%	20%	53.3%	1.5%	2.3%	2.1%	1.106	0.27
TRD	C	70	62%	23%	A	69	47%	23%	B	90	65%	18%	58.0%	1.4%	−4.7%	2.1%	−2.219	0.026
															2.3%	2.1%	1.137	0.26

**Table 1.** First attempt correct rates for each treatment group on the problems they attempted are shown in the "mean" columns for each unit; rates averaged over all three units are shown in the "Total" section. The "Differences" section shows the difference between each row and the one above it, except for the last row, which shows the difference between the DD and the TRD treatment.

### Experiment 2: Pre-Post test Transfer

#### *Introduction:*

The most straightforward method to evaluate overall learning in a course is to administer the same test to students before and after instruction (generally at the beginning and end of the course, as in this study.) This study is designed to answer the question of whether exposure to the problems in the pre-test will enhance students' performance on those same problems on the post test. This is a particularly germane question in a MOOC, because unlike in an on-campus paper test, students are informed whether their answer is correct and given multiple attempts to get it right, hence most (>90%) answer correctly and benefit from the positive feedback.

The average percentage correct for each group on each week is shown in Table 1.

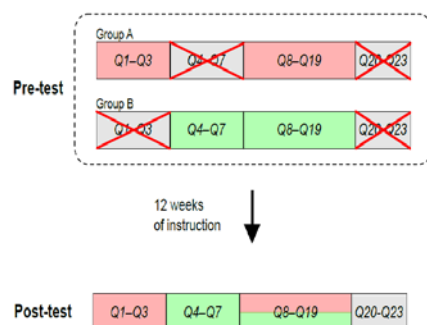
The data suggest that TRD instruction was more effective than deliberate practice in the MC format ( $p=0.026$ ). No conclusive statement can be made based on these data about the relative efficacy of deliberate practice in the DD format vs. the other instructional methods.

#### *Study Setup:*

Figure 1 shows the balanced design for the pre/post-test transfer experiment. The idea is that the two randomly selected groups, A and B, receive identical post-tests, but each receives a pretest with a different subset of post-test problems. If students do learn and transfer from the problems they did on the pre-test, then we should observe that user group A would have an advantage on the set of items unique to pre-test version A and vice versa.

The post-test contained fifteen problems with 23 separate questions. 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 items, Q1–Q3) appeared only on pre-test version A. Two problems (4 edX items, Q4–Q7) appeared only on pre-test version B. Nine problems (12 edX input fields, Q8–Q19) appeared on both pre-test versions A and B. Two problems (4 edX input fields, Q20–Q23) were unique to the post-test.

Users were allowed multiple attempts (usually 2–4, each with right/wrong feedback) on the pre/post-test questions and were not penalized for using multiple attempts. The pre-test was permanently hidden from students after the second week of instruction.



**Figure 1.** Setup for the pre/post-test memory experiment.

### Analysis and Discussion:

A total of 516 users attempted 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. Analysis is complicated by the fact that the A group scores higher than the B group even on the problems that only the B group saw on their pre-test. Correcting for this reduces the apparent superiority of the A group over the B group on the problems that only the A group saw on their pretest to statistical insignificance - except on the second problem where the B groups scores significantly lower than the A group, and lower than the significantly less skillful N-group (that didn't take the pretest). This suggests that the low score of the B group on the second item is a statistical anomaly. Thus it appears that there is no significant evidence for a memory effect on the post test.

Therefore our experiment shows no significant enhancement of post-test scores due to the fact that students were given multiple attempts to obtain the correct answer to the same questions on the pretest. This is good news for those who administer pre-post testing in a MOOC environment and wish to argue that this procedure achieves comparable results to in-class on-paper testing in which the answer is not divulged on either pre or post test.

This is a work in progress, with more in depth data-analysis being performed for these experiments as well as five other experiments. Additional A/B experiments are currently being conducted in our other MOOC.

### Acknowledgements:

We are grateful to Google, MIT, and NSF for supporting our research (but not our conclusions).

### References

- [1] "edX Documentation: Creating Content Experiments." [Online]. Available: [http://edx.readthedocs.org/projects/edx-partner-course-staff/en/latest/content\\_experiments/index.html](http://edx.readthedocs.org/projects/edx-partner-course-staff/en/latest/content_experiments/index.html).
- [2] K. A. Ericsson, R. T. Krampe, C. Tesch-romer, C. Ashworth, G. Carey, J. Grassia, R. Hastie, S. Heizmann, R. Kellogg, R. Levin, C. Lewis, W. Oliver, P. Poison, R. Rehder, K. Schlesinger, and V. Schneider, "The Role of Deliberate Practice in the Acquisition of Expert Performance," *Psychol. Rev.*, vol. 100, no. 3, pp. 363–406, 1993.
- [3] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cogn. Sci.*, vol. 12, no. 2, pp. 257–285, Jun. 1988.
- [4] J. Sweller, P. Ayres, and S. Kalyuga, *Cognitive Load Theory*. New York, NY: Springer New York, 2011.
- [5] A. Pawl, A. Barrantes, D. E. Pritchard, M. Sabella, C. Henderson, and C. Singh, "Modeling Applied to Problem Solving," 2009, pp. 51–54.





# Learning Experiments using AB Testing at Scale in a Physics MOOC

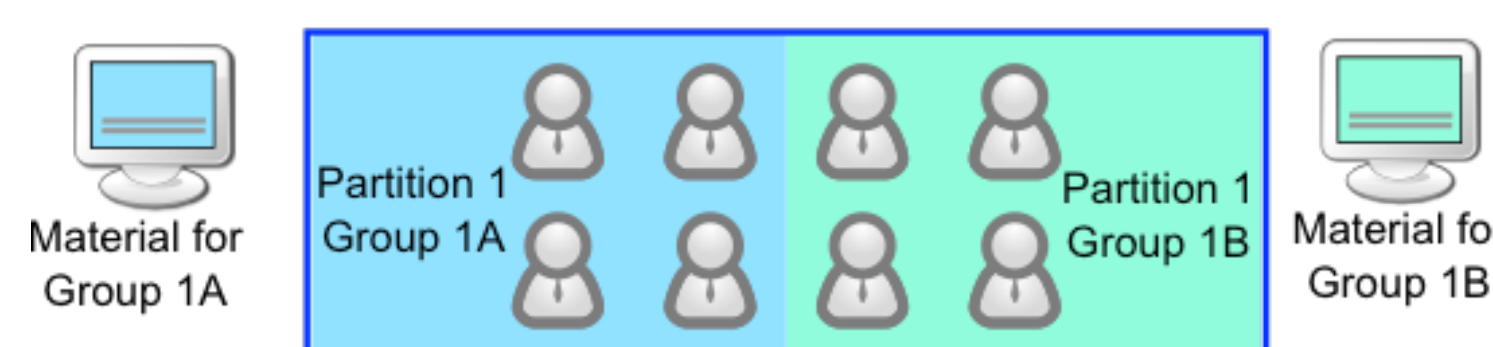
Christopher Chudzicki, Zhongzhou Chen, Youn-Jeng Choi, Qian Zhou, Giora Alexandron David E. Pritchard

**Abstract:** We report results from three treatment/control learning experiments conducted in 8.MReVx: Mechanics Review, a massive open online course (MOOC) offered through edX during summer 2014. Some of our findings include:

- Exposure to items on a MOOC pre-test with multiple attempts and feedback does not affect performance on identical post-test problems. This helps validate the pre-/post-test design in MOOCs
- Adding a diagram to a problem slightly increases correctness and decreases the fraction of students who draw their own in answering the problem.
- Traditional homework problems may be better preparation for traditional assessment than the deliberate practice activities we designed to train individual physics skills.

## The 8MReVx: Mechanics Review MOOC

- Designed for users with some existing knowledge of Newtonian Mechanics
- Uses pedagogical approach Modeling Applied to Problem Solving
- 12 weeks of required material
- Pre-test at beginning, post-test at ending
- Quiz and homework due most weeks; users with at least 60% successful completion earn certificates
- 15,000 users enrolled, 1132 attempted second Homework, 502 earned certificates
- 8MReVx contains several treatment/control experiments:



- Users are randomly assigned to different groups
- Assignment is different for different experiments

**Acknowledgements:** This research was supported by MIT and by a Google Faculty Award

## References:

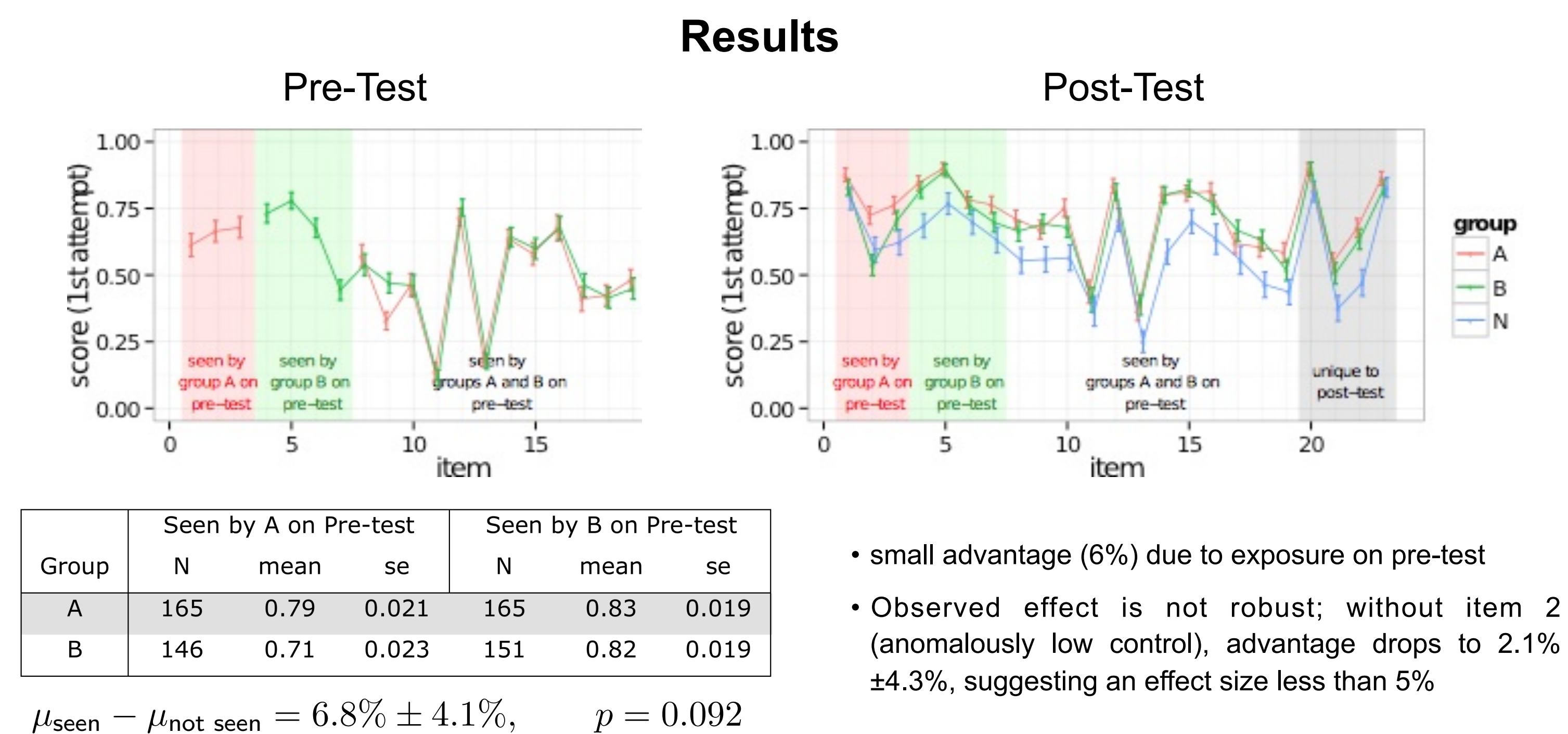
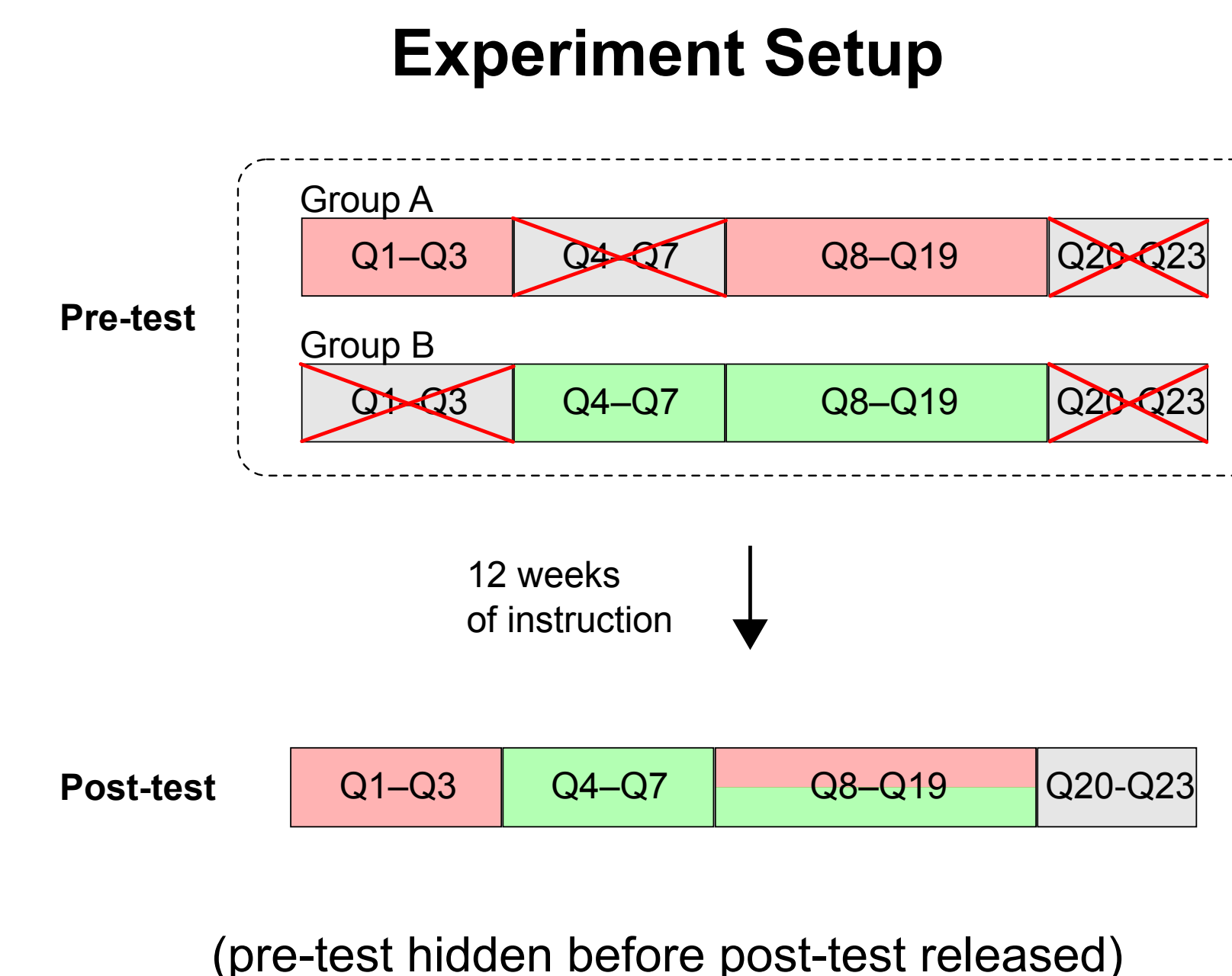
1. K. A. Ericsson, et. al. "The Role of Deliberate Practice in the Acquisition of Expert Performance," Psychol. Rev., vol. 100, no. 3, pp. 363–406, 1993
2. J. Sweller, P. Ayres, and S. Kalyuga, *Cognitive Load Theory*. New York, NY: Springer New York, 2011.
3. A. Pawl, A. Barrantes, D. E. Pritchard, "Modeling Applied to Problem Solving" in *PERC Proceedings 2009*, ed. M. Sabella, C. Henderson, and C. Singh, 2009, pp. 51–54.

## Do MOOC Students Learn During the Pre-test?

We use pre-test / post-test to measure overall learning in our MOOC. Different from classroom pre-/post-testing:

- Users are allowed multiple attempts on each item
- Users receive correct / incorrect feedback on each attempt

**Question:** Does exposure to an item on the pre-test affect your chance of correctly answering the item on post-test?

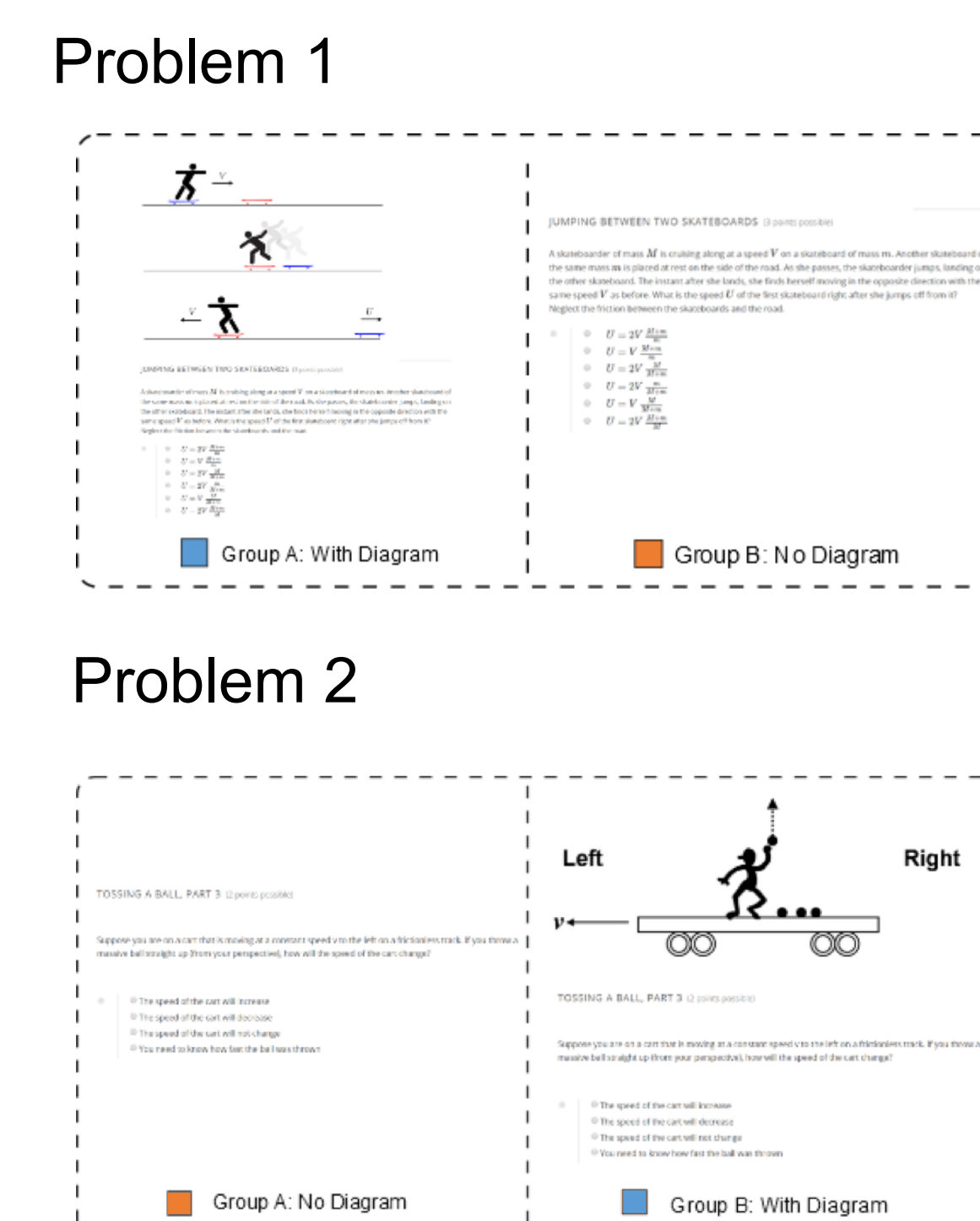


## Should we provide a diagram with each physics problem?

**Questions:** Shall we accompany each physics problem with a diagram?

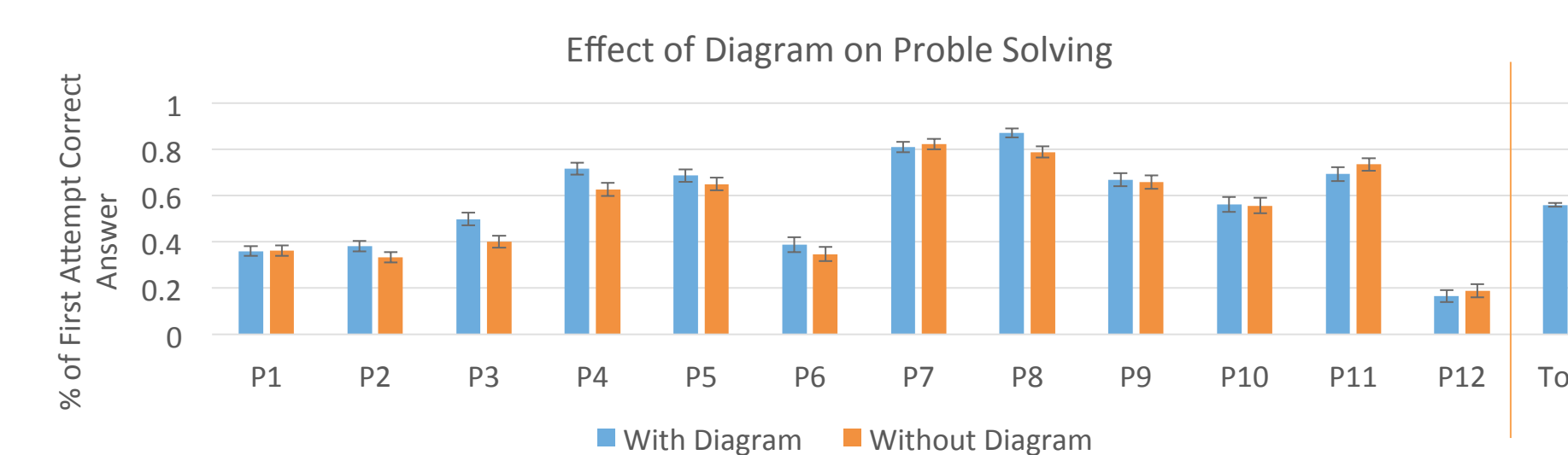
- Does giving diagram assist problem solving?
- Does no-diagram encourage drawing a diagram?

**Method:** In each experiment, Group A gets Problem 1 with diagram, Group B without diagram, and the opposite on Problem 2. In total, 6 pairs of problems (12 problems) were used in this experiment.

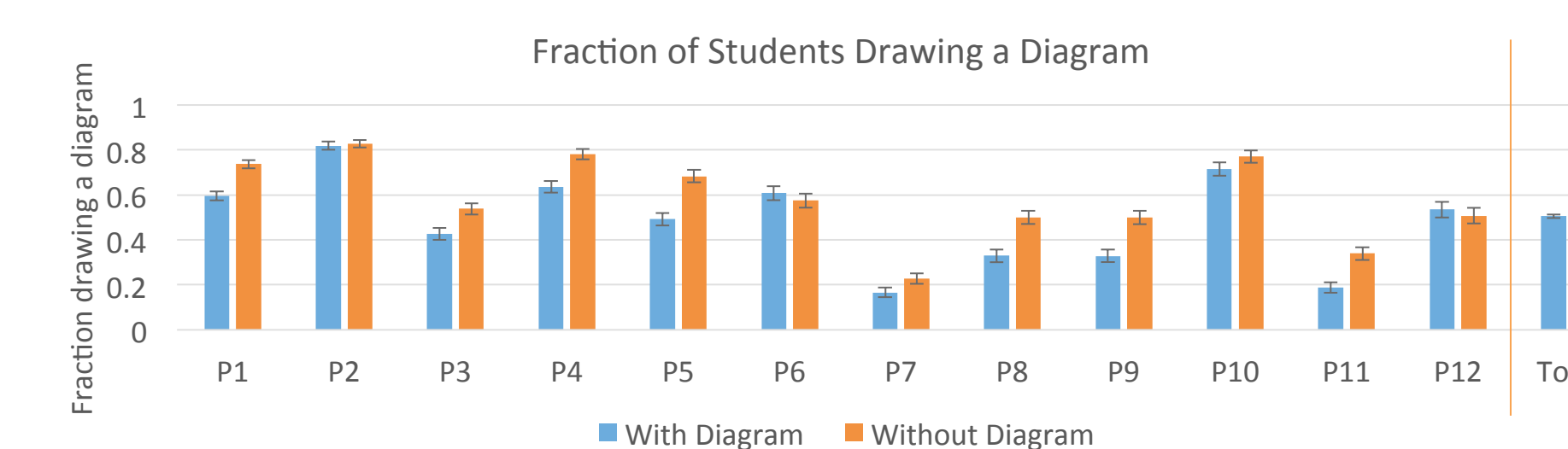


## Results

Giving diagram slightly improves problem solving

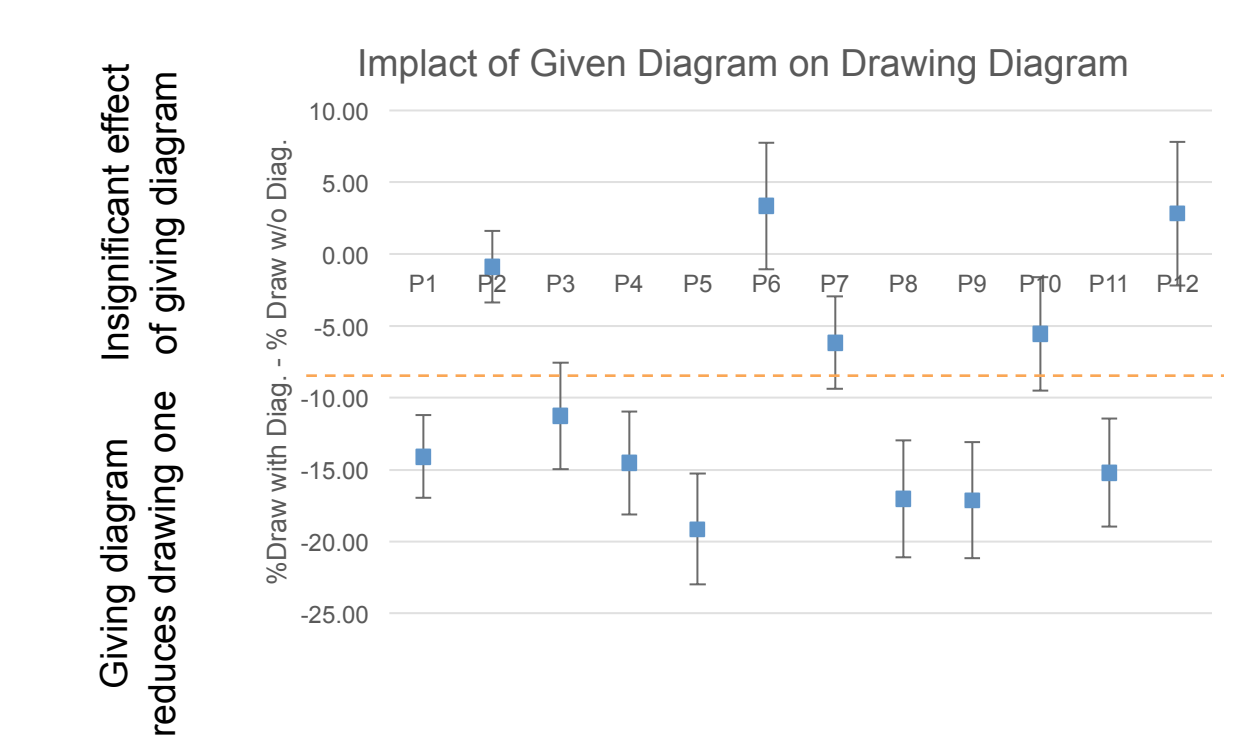


No diagram in general encourages drawing a diagram



Giving a diagram reduces fraction electing to draw a diagram from 0.60 +- x to 0.51+-y. ( $p = 0.0017?$ )

Giving a diagram improves fraction of correct on first try from 0.51 +- x to 0.56+-y. ( $p = 0.013?$ )



## Deliberate Practice vs. Traditional Problems

Traditional physics homework problems often require simultaneous execution of many skills. Previous work by Ericsson suggests that expertise is acquired through *deliberate practice activities (DPAs)*, characterized by:

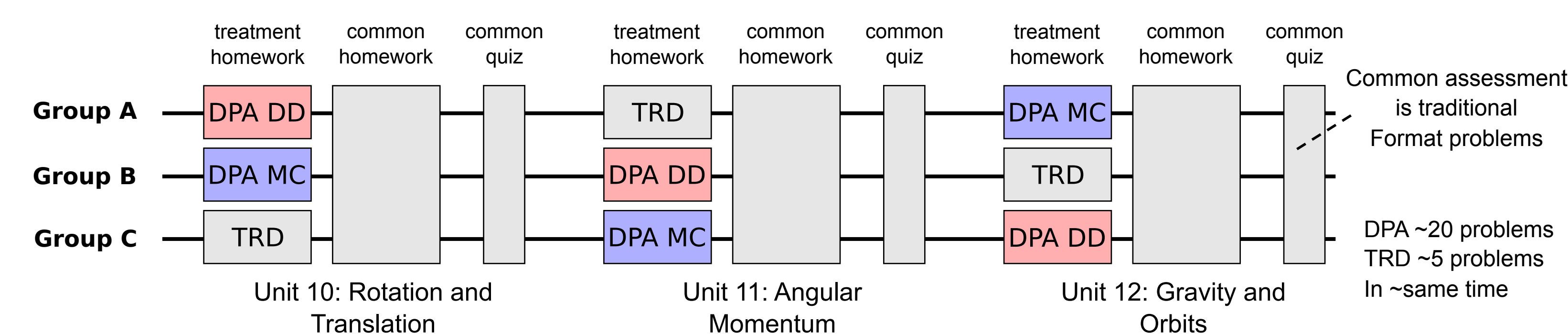
- Breaking up a task into multiple measureable sub-skills
- Focus on improving one skill at a time.
- Provide enough feedback and opportunities to improve

DPAs should be most effective if designed to help students focus on the salient aspects of each activity, i.e., if designed to have low *extraneous cognitive load*.

**Question 1:** Can we use idea of deliberate practice to efficiently build expertise in physics?

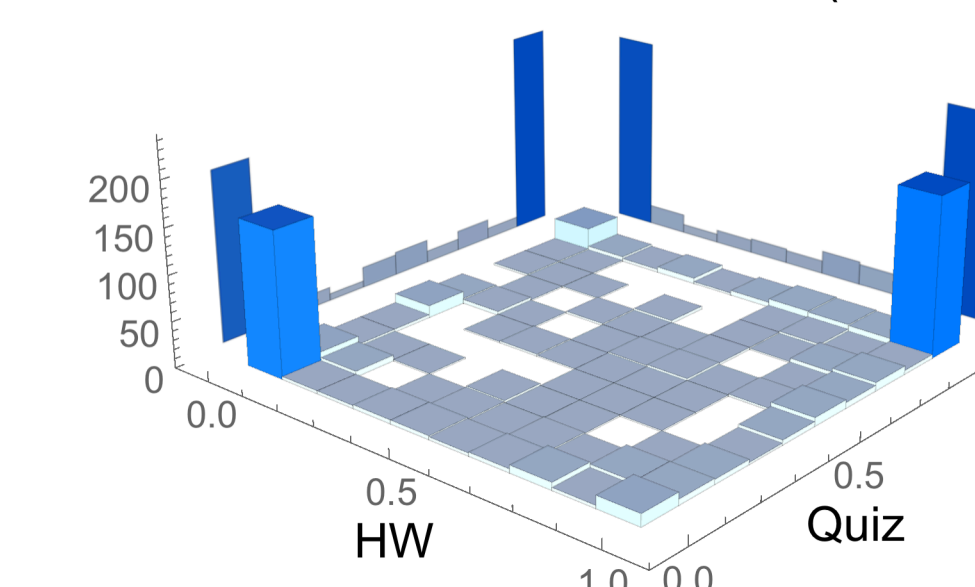
**Question 2:** Can drag-and-drop problem format be used to improve effectiveness of DPAs by reducing extraneous cognitive load?

## Experiment Setup



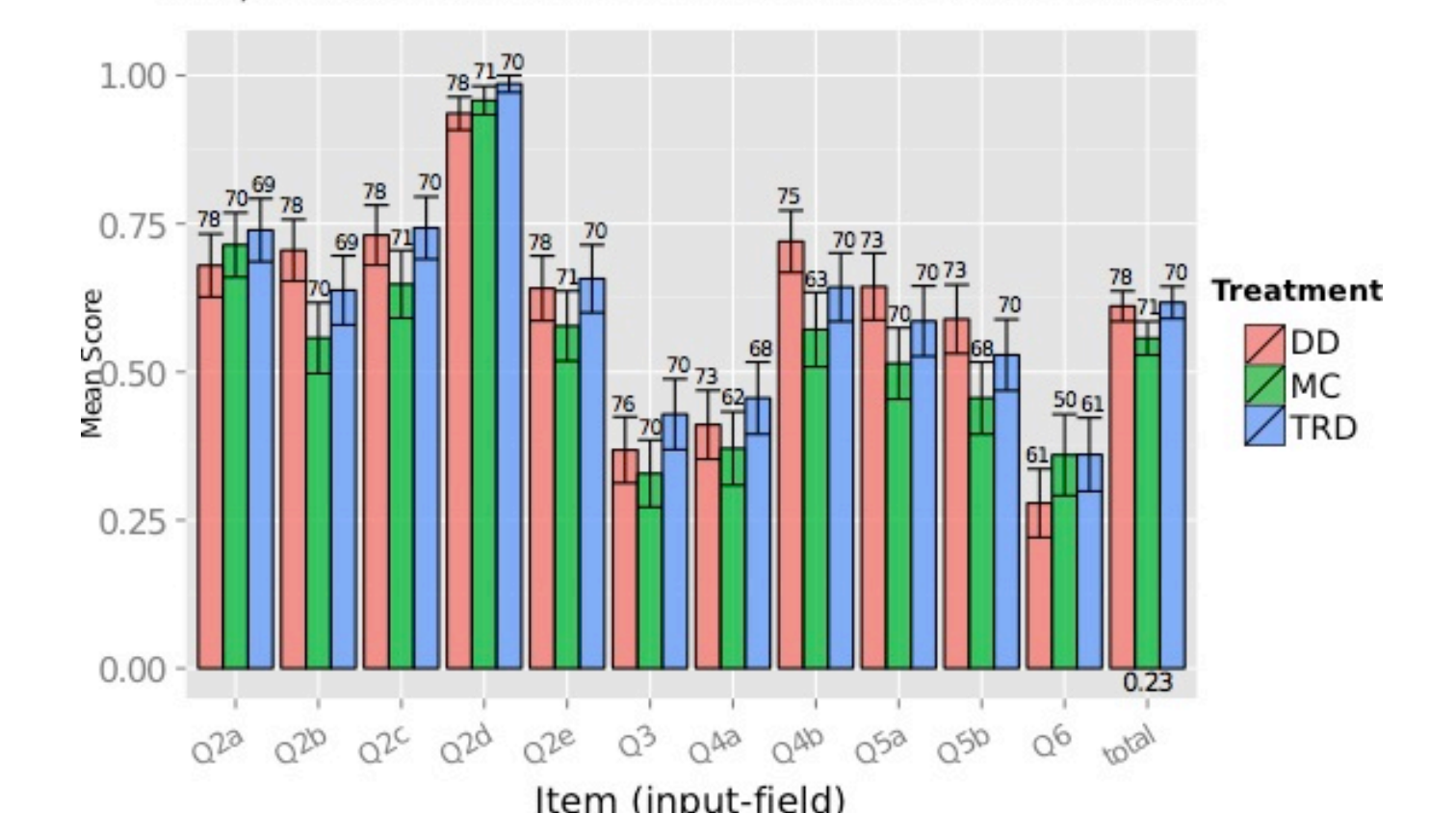
## Results

Completion Rates for Unit 10  
Treatment-HW and Quiz (N=614)



## Quiz 10 Score Analysis (FirstAttempt)

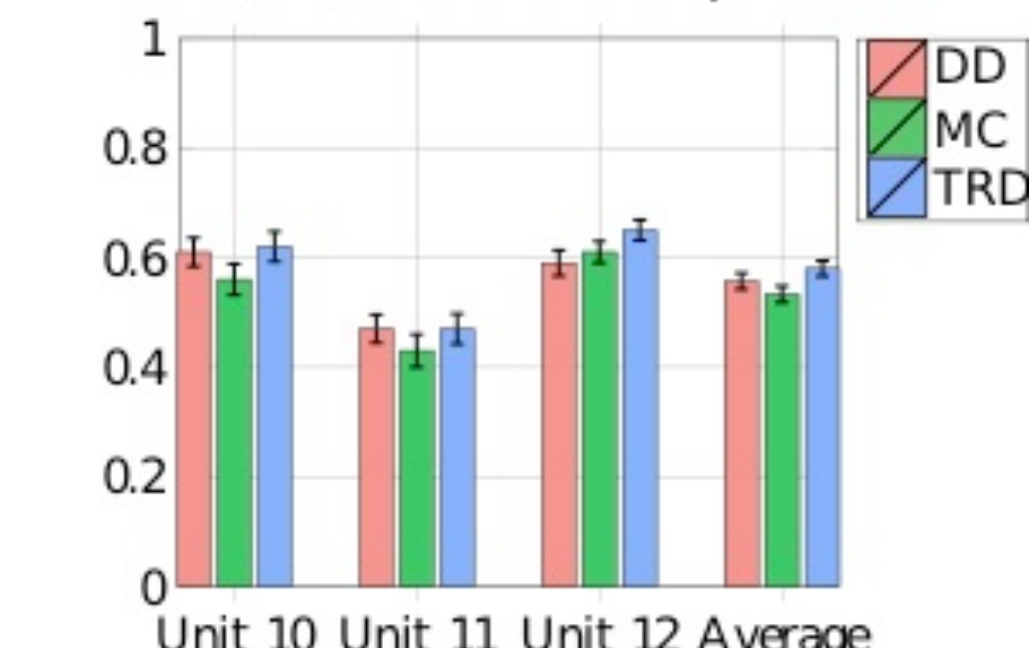
Completion Cut-offs: at least 70% of HW and 70% of Quiz



## Quiz Score Analysis

Treatment	Total Across Units		Differences			
	mean	se	mean	sd	Z-score	p-value
DD (N=237)	55.7%	1.5%				
MC (N=238)	53.3%	1.5%	2.3%	2.1%	1.106	0.27
TRD (N=229)	58.0%	1.4%	-4.7%	2.1%	-2.219	0.026
			2.3%	2.1%	1.137	0.26

## Total Quiz Scores (first-attempt-correct)



**Conclusions:** The group receiving ~ 5 traditional problems performed better than the group receiving ~ 20 DPA-MC exercises on the traditional-format assessment at  $p=0.026$ . No significant difference in assessment scores between DPA-DD and the other two groups was detected.