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.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1145/2724660.2728703">http://dx.doi.org/10.1145/2724660.2728703</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Association for Computing Machinery (ACM)</td>
</tr>
<tr>
<td>Version</td>
<td>Author's final manuscript</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/99202">http://hdl.handle.net/1721.1/99202</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution-Noncommercial-Share Alike</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc-sa/4.0/">http://creativecommons.org/licenses/by-nc-sa/4.0/</a></td>
</tr>
</tbody>
</table>
Learning Experiments Using AB Testing at Scale

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...
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).
**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.

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.

<table>
<thead>
<tr>
<th>Unit 10</th>
<th>Unit 11</th>
<th>Unit 12</th>
<th>Total</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment Group</td>
<td>N</td>
<td>mean</td>
<td>sd</td>
<td>N</td>
</tr>
<tr>
<td>DD</td>
<td>A</td>
<td>78</td>
<td>61%</td>
<td>23%</td>
</tr>
<tr>
<td>MC</td>
<td>B</td>
<td>71</td>
<td>56%</td>
<td>24%</td>
</tr>
<tr>
<td>TRD</td>
<td>C</td>
<td>70</td>
<td>62%</td>
<td>23%</td>
</tr>
</tbody>
</table>

*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.

**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 pre-test 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
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 8MReVx: 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 than traditional-style assessments.
- Adding a diagram may help students focus on the salient aspects of a problem.
- 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 measurable sub-skills.
  - Focus on improving one skill at a time.
  - Provide enough feedback and opportunities to improve.

References:

The 8MReVx: Mechanics Review MOOC

- Designed for users with some existing knowledge of Newtonian Mechanics.
- Uses pedagogical approach Modelled 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.

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.

Experiment Setup

Problem 1

1. Question: Does no-diagram encourage drawing a diagram?

2. Problem 2

Results

Effect of Diagram on Problem Solving

(a) No diagram in general encourages drawing a diagram

(b) Giving a diagram reduces fraction of incorrect answers

DPA vs. Traditional Assessments

Problem Solving in a Physics MOOC

Different from Deliberate Practice

Focus on improving one skill at a time.

Common assessment of traditional formal problems

DPA >0 problems

Quiz Score Analysis

Quiz 10 Score Analysis (First Attempt)

Completion Rates for Unit 10 Treatment-HR and Quiz (N=16)

Quiz Score Analysis

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.