

Validating the pre/post-test in a MOOC environment

Christopher Chudzicki¹, Zhongzhou Chen¹, Qian Zhou², Giora Alexandron¹, and David E. Pritchard¹

¹*Massachusetts Institute of Technology, Department of Physics*

77 Massachusetts Avenue, Cambridge, MA, USA 02139

²*Qinghua University*

30 Shuangqing Rd, Haidian, Beijing, China, 100084

Abstract. A standard method for measuring learning is to administer the same assessment before and after instruction. This pre/post-test technique is widely used in education research and has been used in our introductory physics MOOC to measure learning. One potential weakness of this paradigm is that post-test performance gains may result from exposure on the pre-test instead of instruction. This possibility is exacerbated in MOOCs where students receive multiple attempts per item, instant correct/incorrect feedback, and unlimited time (until the due date). To find the size of this problem in our recent MOOCs, we split the student population into two groups, each of which received identical post-tests but different subsets of post-test items on their group pre-test. We report a small overall advantage ($2.9\% \pm 1.7\%$) on post-test items due to pre-test exposure. However, this advantage is not robust and is strongly diminished when one obviously anomalous item is removed.

PACS: 01.40.Fk, 01.40.gf

I. INTRODUCTION

Measuring the learning that occurs within a massive open online course (MOOC) is an important step toward establishing the effectiveness of this new learning environment. A standard method to evaluate the overall learning that occurs during a course is to administer the same assessment before and after instruction. Often, a calibrated assessment such as the Force Concept Inventory (FCI) [1] or Mechanics Baseline Test (MBT) [2] is used. Recently, this pre-/post-test technique was used in our introductory physics MOOC to measure learning and show that equal learning occurred among different cohorts of certificate earners [3].

A common concern about the pre-/post-test design has been that exposure to identical items on the pre-test could affect student scores on the post-test. Data from the University of Minnesota [4] suggest that this is not the case for the FCI given in traditional, on-campus courses. However, the concern that post-test scores are affected by exposure to identical pre-test items—i.e., that memory learning occurs during the pre-test—is exacerbated by several factors in the MOOC setting. For example, in our MOOCs on the edX.org platform: (1) users are given multiple attempts on each test item; (2) users are given correct/incorrect feedback on each attempt; (3) the pre- and post-tests are "open", i.e., users can take advantage of outside materials and in-course materials; and (4) users have essentially unlimited time (several weeks) to complete the pre- and post-tests.

In this paper, we report results from a randomized control study run in two MOOCs in order to measure how

exposure to pre-test items affects student performance on a post-test.

II. METHODS

The study was conducted in the MOOCs 8.MReVx: Mechanics ReView and 8.MechCx: Advanced Introductory Classical Mechanics. Both MOOCs were introductory, college-level, calculus-based MOOCs run on the edX platform. The course began with an un-graded, pre-test and culminated in a graded post-test. Students were given several attempts on most problems within the course, including the pre-test and post-test problems. Correct/incorrect feedback was given on each problem. Students could not view the correct answer to pre-test problems and the pre-test was hidden to students after the first few weeks of each course. Table 1 shows overall course statistics.

TABLE 1. Course Statistics

	8.MReVx (Summer 2014)	8.MechCx (Spring 2015)
Length (weeks)	12 required 2 optional	15 required 1 optional
Participants ¹	4533	2,685
Certificate Earners	502	199
Attempted pre-test	2,483	1,184
Attempted post-test	516	239

¹*Participants* includes all users who ever attempted a problem within the course, excluding beta testers.

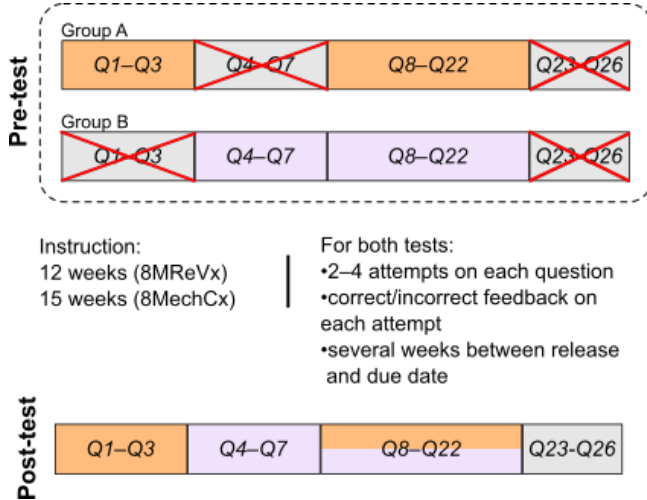


FIG 1. Each group received identical post-tests, but a different subset of post-test items on their group pre-test.

Study Setup

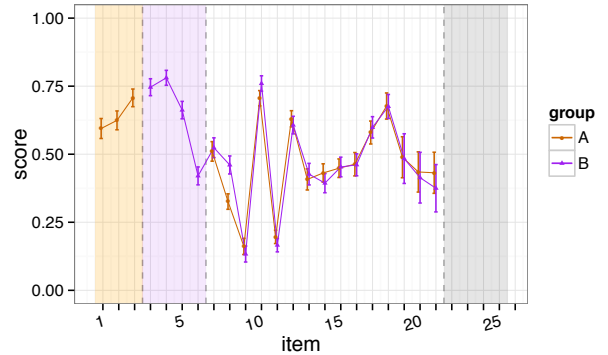
Each student who elected to access the pre-test was randomly assigned [5] to one of two groups, A and B. Both groups received identical post-tests at the end of the course, but were given a different subset of post-test items on the ungraded pre-test, as depicted in Fig. 1.

During both 8MReV (2014) and 8MechCx (2015), the post-test contained 15 physics problems (some multipart) consisting of a total of 23 separate edX questions, or *items*. The 8MReVx and 8MechCx post-tests were identical except for three items in 8MReVx that were replaced by three other items in 8MechCx. Three questions appeared only on pre-test version A (items Q1-Q3); two problems appeared only on pre-test version B (items Q4-Q7); questions Q7-Q22 appeared on both pre-test versions (with Q16-Q18 only in 8MReVx and Q19-Q22 only in 8MechCx); and four problems appeared only on the post-test (Q23-Q26).

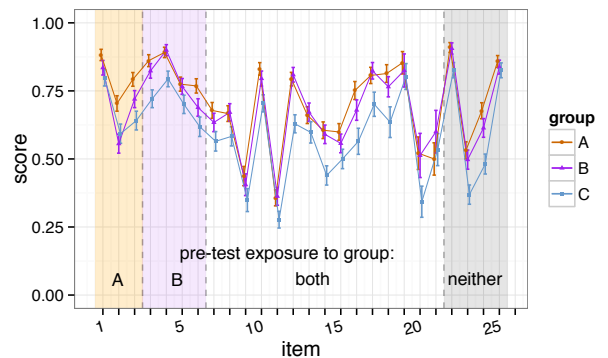
III. RESULTS AND DISCUSSION

A total of 755 users (excluding beta-testers) attempted the post-test in 8MReVx/8MechCx; not all users attempted all problems on the post-test. On average, users who attempted the post-test tried 85% of items on the post-test. Each user was assigned to either version A or version B of the pre-test, but not all of these 755 users attempted the pre-test. Users who did attempt the pre-test attempted 89% of items on average.

To avoid score saturation, we use the first-attempt-correct rate to quantify performance on the pre- and post-tests. Figure 2 shows the mean first-attempt-correct rates for 8.MReVx and 8.MechCx users on each pre- and post-test item for three groups of users: groups A ($N_A=277$) and B ($N_B=236$) include users who were assigned to versions A and B of the pre-test and attempted at least some pre-test



(a) pre-test first-attempt-correct rates by group



(b) post-test first-attempt-correct rates by group

FIG 2. Pre-test and post-test first-attempt-correct rates per item for each user group; error bars show one standard deviation. Dashed vertical lines separate item categories.

TABLE 2. Post-test performance by group and pre-test exposure

Group	Seen on pre-test by	N	mean	sd	sem
A	A	165	0.79	0.27	0.021
A	B	165	0.83	0.24	0.019
A	A and B	112	0.71	0.17	0.016
A	neither	169	0.75	0.25	0.019
B	A	146	0.71	0.28	0.023
B	B	151	0.82	0.23	0.019
B	A and B	106	0.71	0.19	0.019
B	neither	146	0.72	0.25	0.021
C	A	95	0.68	0.33	0.034
C	B	92	0.71	0.31	0.032
C	A and B	56	0.62	0.23	0.031
C	neither	90	0.66	0.28	0.029

problems; group C ($N_C=241$) includes users who did not attempt any pre-test problems.

The performance of each user group (A, B, C) on each of these item categories (seen by A on pre-test, seen by B on pre-test, seen by A and B on pre-test, seen by neither group on pre-test) is summarized in Table 2. The mean first-attempt-correct rate for each user group on each item

category is calculated only over that subset of users who attempted all items in that particular item category.

A. Non-Participation and Group Size

The most striking feature of Figure 2 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 C). This is true across all item categories, including items that were unique to the post-test, suggesting 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 students who are weaker or less interested in learning (group C) elect not to participate in the optional pre-test.

The size of group C indicates one difficulty of administering an ungraded pre-test in a MOOC: whereas Henderson reported that only 3.2% of residential students showed a lack of serious effort on proctored, ungraded, classroom administrations of the FCI pre-test, we find that 32% of our users who attempted the graded post-test did not attempt the optional pre-test.

Additionally, we comment that the difference size between groups A and B is larger than one would expect by random assignment. We are investigating possible causes for this difference, and suggest that the altered order of pre-test items between groups A and B may have affected pre-test participation rates.

B. Advantage from Pre-test Exposure

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 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, suggesting that students in group A are somewhat more skilled than students in group B.

To quantify the post-test advantage provided to users by exposure to identical items on the pre-test, we average the advantage to A on items seen by A and the advantage to B on items seen by B.

TABLE 3. Differences in post-test performance by group and pre-test exposure

Difference	Pre-test exposure	mean	sem	Z-score	<i>p</i>
A-B	A	0.072	0.027	2.7	0.007
B-A	B	-0.013	0.022	-0.59	0.56
Average		0.029	0.017	1.7	0.089

TABLE 4. Differences in post-test performance by group and pre-test exposure with i2 removed.

Difference	Pre-test exposure	mean	sem	Z-score	<i>p</i>
A-B	A	0.046	0.029	1.49	0.14
B-A	B	-0.013	0.022	-0.59	0.56
Average		0.015	0.018	0.84	0.40

Although group A's advantage on problems seen by group A on the pre-test is significant at the $p=0.007$ level, this advantage is likely due to a combination of increased skill and prior exposure. The average of the advantages cancels out the better overall ability of group A, revealing the advantage due to prior exposure. The advantage due to prior exposure is insignificant ($p=0.089$) with a mean score shift of $2.9\% \pm 1.7\%$. In comparison, the overall normalized gain [6] on problems attempted on both the pre- and post-test by users in groups A and B was $32\% \pm 2.5\%$. Correcting for prior exposure advantage suggests a normalized gain closer to $29\% \pm 3.0\%$.

One item on the post-test, i2, showed an especially large difference in performance between groups A and B. 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. This item is one of three "problem decomposition" activities in the entire course, and so it seems somewhat plausible that increased exposure to this rare problem format on the pre-test could reduce group B's later performance relative to group A. However, group B's score on this item is also anomalously lower than group C (no pretest) even though they usually outperform group C by $8.8\% \pm 1.0\%$. This suggests that the low performance of group B on i2 is a statistical anomaly. Removing this single item from our analysis reduces all group differences to statistical insignificant, and dramatically reduces the advantage due to prior exposure (Table 4).

IV. CONCLUSIONS AND FUTURE DIRECTIONS

Overall, our study showed little evidence for enhancement of post-test performance from exposure to the same items on the pre-test, even in a MOOC setting where multiple attempts are allowed and correct/incorrect feedback is given on pre-test items. This bodes well for those who plan to use pre-/post-testing to measure learning in MOOCs even though the time is unlimited and the several graded attempts are allowed.

One advantage of MOOC studies such as this is that they take relatively little effort to alter and/or repeat. In future iterations of this experiment, we hope to solidify our findings by increasing the number of post-test items unique to each pre-test group.

ACKNOWLEDGEMENTS

We are grateful to Google, MIT, and NSF for supporting our research.

- [1] D. Hestenes, M. Wells, and G. Swackhamer, Phys. Teach. **30**, 141 (1992).
- [2] D. Hestenes and M. Wells, Phys. Teach. **30**, 159 (1992).
- [3] K. F. Colvin, J. Champaign, A. Liu, Q. Zhou, C. Fredericks, and D. E. Pritchard, Int. Rev. Res. Open Distance Learn. **15**, (2014).
- [4] C. Henderson, Phys. Teach. **40**, (2002).
- [5] http://edx-partner-course-staff.readthedocs.org/en/latest/content_experiments/content_experiments_overview.html, retrieved 6/30/2015
- [6] R. Hake, Am. J. Phys. **1** (1998).