IRRODL
THE INTERNATIONAL REVIEW OF RESEARCH IN OPEN AND DISTANCE LEARNING

Learning in an Introductory Physics MOOC: All Cohorts Learn Equally, Including an On-Campus Class





Kimberly F Colvin[1], John Champaign[1], Alwina Liu[1] (not shown), Qian Zhou[2], Colin Fredericks[3], and David E Pritchard[1]
[1]Massachusetts Institute of Technology, USA, [2]Tsinghua University, China, [3]Harvard University, USA

Abstract

We studied student learning in the MOOC 8.MReV Mechanics ReView, run on the edX.org open source platform. We studied learning in two ways. We administered 13 conceptual questions both before and after instruction, analyzing the results using standard techniques for pre- and posttesting. We also analyzed each week's homework and test questions in the MOOC, including the pre- and posttests, using item response theory (IRT). This determined both an average ability and a relative improvement in ability over the course. The pre- and posttesting showed substantial learning: The students had a normalized gain slightly higher than typical values for a traditional course, but significantly lower than typical values for courses using interactive engagement pedagogy. Importantly, both the normalized gain and the IRT analysis of pre- and posttests showed that learning was the same for different cohorts selected on various criteria: level of education, preparation in math and physics, and overall ability in the course. We found a small positive correlation between relative improvement and prior educational attainment. We also compared homework performance of MIT freshmen taking a reformed on-campus course with the 8.MReV students, finding them to be considerably less skillful than the 8.MReV students.

Keywords: MOOC; edX; item response theory; learning gain

Introduction

The recent release of hundreds of free online courses in MOOCs (massive open online courses) by organizations such as Coursera, edX, and Udacity has been so dramatic that an article in the New York Times proclaimed 2012 the "Year of the MOOC" (Pappano, 2012). These MOOCs, often digitizations of standard, relatively introductory courses from top 50 universities (and especially MIT, Harvard, Berkeley, and Stanford), have provoked multidimensional discussions and special issues of various publications. Nevertheless, few studies have attempted to use MOOC data to address the central question: "is there learning in MOOCs?" Even though documenting learning is a stated goal of some institutions offering MOOCs, there have been few developments in answering this question (Hollands & Tirthali, 2014). In their thorough treatment of the current state of MOOCs and related research, Hollands and Tirthali point out that the lack of consistent data and the voluntary nature of the student participants, who aren't forced to take a pre- and posttest, for example, has been a roadblock. Hollands and Tirthali did find research on student retention, motivation, and behaviors within a MOOC.

In this paper, we report an initial study of learning in a MOOC, 8.MReV – Mechanics ReView – offered from June 1 to August 27, 2013 on the open source platform edX.org. The course materials were written by the RELATE education group (Research in Learning, Assessing, and Tutoring Effectively, http://RELATE.MIT.edu). This is a "second course" in introductory Newtonian Mechanics, designed to help students familiar with the topic at a high school level gain a more expert-like perspective on the subject by learning a categorization scheme for the domain and applying it to work through sophisticated problems that typically involve several physics principles simultaneously (e.g., conservation of both momentum and energy). In addition, we made a concerted effort to attract high school physics teachers to enroll in our course.

We emphasize that measurements of learning in a MOOC are made in an online environment that allows students to consult reference materials both inside the course (e.g., course resources available to students at that point in time) and outside the course (e.g., Google, Wikipedia, or a textbook). This applies to homework as well as to the pre- and posttesting, in contrast to on-campus pre-post testing which is done in a closed-book, no Internet environment. Furthermore, because on-campus assessments are done on paper, students are usually restricted to only one response. In contrast, online students are usually allowed several attempts to get the answer correct and are told whether each response is correct. (We only analyzed the first attempt to make it more similar to a traditional pre- and posttest.) Thus, our definition of learning involves improvement in answering questions *with*, rather than *without*, outside assistance. While this may be a more authentic activity than closed-book examinations, it blurs the comparison of our pre- to posttest results with those from on-campus students.

We used two major approaches to evaluate learning in our MOOC. The first was to give an identical pretest and posttest using the same set of mostly conceptual questions

(Hestenes, Wells, & Swackhamer, 1992). The results were analyzed in terms of the fractional reduction in the number of incorrect answers on the pretest as measured by the posttest. This quantity is referred to as the normalized gain by Hake (1998).

The second approach involved using item response theory (IRT) to analyze the pre- and posttest results as well as the weekly performance of the students. IRT establishes an "ability" for each student based not on total score, but on the difficulty and discrimination of the questions (items) that that student attempted (Meyer & Zhu, 2013). (Discrimination, or slope, is related to the difference in performance of high versus low ability students on that item.) This is especially important in MOOCs where not all students respond to the same set or number of items. We selected cohorts based on education level, preparation in calculus and physics, and on overall skill in the course. The weekly performance of selected cohorts of MOOC students were compared with each other as well as with students in an on-campus course (8.011). The 8.011 IRT analysis was based on homework assignments containing roughly two-thirds of the same questions as the MOOC, also delivered on the same edX platform.

Data

Description of Mechanics ReView MOOC: 8.MReV

The 8.MReV course grew from a short Mechanics ReView course that runs at MIT during January for students who received a D in MIT's large-enrollment fall Mechanics course. The key feature of the ReView course was that faculty and staff interact with two-person groups of students to help them focus on problem solving using our modeling applied to problem solving pedagogy (Pawl, Barrantes, & Pritchard, 2009). This in-class work required preparing the students for class, a need that RELATE met by developing an online eText and assigning pre-class homework at different levels of difficulty. These online materials, augmented by additional problems and weekly quizzes, were offered as a free open online course twice in 2012 using the LON-CAPA platform. Development of the 8.MReV course studied here involved transferring much of this course to the edX platform and supplementing it with more problems (Fredericks et al., 2013). 8.MReV was run in the summer of 2013 with both general and teacher-targeted publicity.

8.MReV includes three basic types of problems: (1) checkpoint questions embedded in the eText for the purpose of guiding the reader and checking for understanding, (2) homework problems at different levels of difficulty (Teodorescu, Pawl, Rayyan, Barrantes, & Pritchard, 2010), and (3) weekly test questions. In the course, the tests are referred to as quizzes, in part, to lower student anxiety. Because there is no final exam the quizzes count heavily toward earning the certificate. Students must obtain at least

60% of the total credit available to earn a certificate; 1,030 students earned certificates in the summer 2013 8.MReV.

The online course included 288 homework items and 115 quiz items. These common questions allowed us to easily compare performance in the two courses. There were three optional units at the end of 8.MReV; because these units were not required to earn a certificate they were not included in this analysis. Although approximately 17,000 people signed-up for 8.MReV, most dropped out with no sign of commitment to the course; only 1500 students were "passing" or on-track to earn a certificate after the second assignment. For the IRT analysis we included only the 1,080 students who attempted more than 50% of the questions in the course, 95% of whom earned certificates. Most of those completing less than 50% of the homework and quiz problems dropped out during the course and did not take the posttest, so their learning could not be measured.

For most homework and quiz items students were allowed multiple attempts at a correct answer: several for multiple-choice items and typically ten attempts for symbolic, free-response items. Informing a student of an incorrect response and allowing additional attempts improves test information and affords a more reliable ability estimate than only using the student's first response (Attali, 2010). We only modeled up to eight attempts, since very few students used more then eight attempts for quiz and homework items. Most items only needed three or four attempts to accurately model student behavior.

Methods and Theoretical Framework

Pre- and Posttesting in the MOOC

A pretest was given before students started working with the materials in 8.MReV. The pretest consisted of 15 questions, three of which came from the Mechanics Baseline Test (Hestenes & Wells, 1992) and four of which came from the Mechanics Reasoning Inventory (Pawl et al., 2011). The posttest contained the same 15 questions plus two multiple-choice items from the Mechanics Baseline Test. See Table 2 in the Appendix for a list of pre- and posttest problems.

The pretest, called Quiz 0, was given at the beginning of the course with this note from the instructor, Professor Pritchard.

Why Quiz 0?

It may seem strange to start a course with a quiz, but it is important for you to take this quiz, especially because this course is for people with "some knowledge of mechanics."

It will give us important insight into what you and the class bring to the course in terms of various skills.

It will give you practice taking a quiz (and will not count in your grade).

It will give you an idea of the variety of problem types you'll experience in this course.

In the future analysis of these data we will be able to give new students guidance on whether they're ready for this course.

So we hope you will make a conscientious effort to do well on this quiz.

This pretest was then hidden from the students several weeks into the course, so that they could neither review these questions nor refer to them when answering the posttest. The posttest questions were contained in the last two weekly tests, although some of the better students had amassed sufficient points for a certificate and didn't take these tests. Of the 3,899 students who attempted at least one item on the pretest, the mean number of items attempted was 10.0; of the 1,117 students who attempted at least one item on the posttest, the mean number of items attempted was 8.2.

These 15 questions weighed conceptual knowledge more heavily than algebraic ability. With the exception of two questions requiring symbolic entries, all of the items were multiple choice questions. Two questions were given only on the posttest as part of a study on the residual effects of student memory on pre- and posttesting, and are not included in this analysis.

Item Response Theory (IRT)

IRT judges student ability by taking into account a student's specific performance on each item. An item is a single question that demands a unique student response that generally can be judged right or wrong. We considered each item separately, even where two items are from the same multi-part problem. IRT stands in contrast to classical test theory where the unit of analysis is the entire test, usually scored as the total number of items correct (Mellenbergh, 2011; Crocker & Algina, 1986).

An advantage of IRT is that it gives accurate estimates of students' abilities even when students have not taken the same set of items. This is particularly important in this study because students do not need to complete all of the homework and quiz questions in 8.MReV or 8.011 to pass the course, so students do not generally attempt all available

problems. IRT is also preferable because it extracts more information than simply using the total number of items correct by accounting for the difficulty of each item and each item's ability to discriminate between students of higher and lower abilities (Hambleton, Swaminathan, & Rogers, 1991).

IRT relates a student's performance on a set of items to the student's ability (skill) on an underlying trait or proficiency, referred to as *s*, in this study. Many IRT models exist; all contain at least one parameter related to the item and at least one parameter related to the student (Hambleton et al., 1991). IRT allows students and items to be placed along the same proficiency scale, where higher numbers indicate more difficult items and more proficient students. IRT's 2-parameter logistic model (2PL) is a common example:

$$P_i(s) = \frac{e^{a_i(s-d_i)}}{1+e^{a_i(s-d_i)}}$$ (Hambleton et al., 1991),

where a_i and d_i are the parameters for item *i* and *s* is the examinee's ability, also referred to as proficiency or ability. $P_i(s)$ is the probability that an examinee with ability *s* will correctly respond to item *i*. The *d*-parameter is the item difficulty parameter. The *a*-parameter is the discrimination parameter and can be thought of as the correlation between performance on an item and performance on the test or complete set of items as a whole. An assumption of IRT is unidimensionality, that there is only one dimension or factor affecting the likelihood of a student's correct response, namely, the student's underlying ability. If unidimensionality holds, the probability of correctly answering an item should increase as the level of ability increases.

IRT is sophisticated "grading with respect to a curve": Student abilities are constrained in an IRT analysis to have a mean of 0 and a standard deviation of 1. Thus a time series of IRT scores of a student in a class that is learning does not show absolute learning (as measured above by pre-post testing), but rather improvement relative to "class average." However, weekly IRT ability is a good measure for comparing two different cohorts undergoing two different pedagogical treatments, or even different cohorts of students undergoing the same pedagogical treatment, for example, to investigate the effects of demographics or study patterns on relative learning rates.

Multiple Attempts

To incorporate multiple attempts (IRT), we modeled student ability with an extension of item response theory that accounts for ordered response categories, not just binary (right or wrong) responses. Samejima's graded response model (1997) is an extension of IRT's 2-parameter-logistic model described earlier and was developed to model ordered scores or responses to an item. This could be an essay scored with a point-based rubric, for example, or the number of correct steps an examinee performs in an algebra problem with a clear set of steps required for a complete response. As suggested by Attali (2010), modeling the number of attempts a student needs before a correct

response is analogous to these ordered categories, with the fewer attempts required indicating more ability.

The graded response model (GRM) models the likelihood that an examinee with a given ability will provide a response in each category. In this study, the categories correspond to the attempt with a correct response. The probability of a correct answer on the second attempt is modeled as the product of the probabilities of correct responses on the third, fourth, and fifth attempts and the probability of an incorrect response on the first attempt. The GRM assumes that a positive response in category *n* implies positive responses in all lower, that is, easier categories. This assumption is reasonable in this application where it would be reasonable to assume that an examinee who correctly responds to an item on the 2nd attempt (and is so informed in real time) would correctly answer on the third, fourth, and fifth attempts if indeed they were made. The item parameters and student ability estimates are calculated using maximum likelihood estimation via the psychometric software MULTILOG (Thissen, 1991).

Item Calibration

Using the graded response model, we initially calibrated quiz and homework items separately. We first looked for items not fitting the model, meaning that for a particular item students with less ability were more likely to respond correctly to the item on an earlier attempt than strong students, for example. These items prevented the model from converging such that it was impossible to calibrate the other items. We identified 7 quiz and 32 homework out of 138 and 256, respectively, for removal from the IRT analysis. Because there were so many items in total, this had little to no effect on the final estimates of student ability. Once a decision was made about which items to remove, all homework and quiz items were calibrated simultaneously using the combined student pool from 8.MReV and 8.011. The item parameters from this joint calibration were then used to obtain ability estimates for each student's weekly homework and quiz performances. We calculated each student's ability on each weekly topic. The distribution of student abilities for each week was re-centered such that the mean ability for each week was zero, allowing a week-by-week comparison of changes in ability.

Pre- and Posttest Results

Pre- and Posttest Analysis with Normalized Gain

To assess whether students exhibited learning in the MOOC, we analyzed the pre- and posttest results in two different ways: using traditional pre- and posttesting procedures (Hake, 1998) involving normalized gain and using IRT. Item response theory can judge

the ability of students from different subsets of the items, allowing us to include a larger fraction of our students in the pre-post comparison. To effectively compare pre- and posttest scores using the normalized gain technique, students need to have attempted the same set of questions on both tests, which limits the number of students in each cohort. In part to increase the sample size, the pre-post test analysis was performed on two subsets of questions: (1) six questions involving force and motion that could be compared with Hake's study (1998), (2) five purely conceptual questions on more advanced topics, and (3) seven questions consisting of these five plus two questions requiring symbolic responses rather than multiple choice. There were 419 students who attempted all six questions in subset 1 on both the pre- and posttests, 343 students who attempted all five subset 2 questions on both the pre- and posttests, and 176 students for subset 3. Data for these three question sets are presented in Hake's format in Figures 1-3 where the various cohorts are analyzed independently.

The 6,000-student study by Hake (1998), which investigated about 60 different classes ranging from high school to top quality colleges, showed that the normalized gain is typically 0.23 for traditionally taught courses, but increases to about 0.48 for interactively taught courses. The clearest comparison with Hake's numbers is subset 1, questions involving force and motion, as shown in Figure 1.

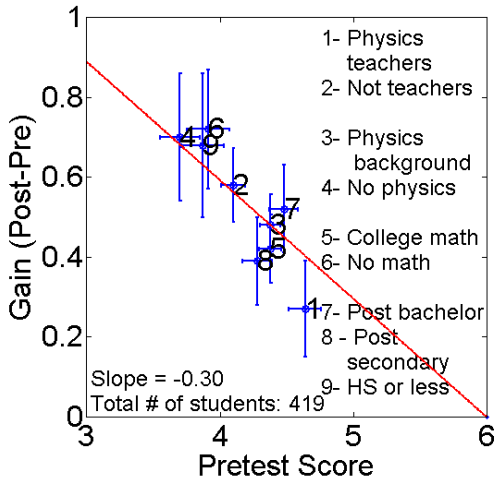

Figure 1. The negative slope of the red line, constrained to go through the point (6,0), indicates the normalized gain that best fits the 419 students who answered all 6 force-related items on both pre and post tests. The mean pretest and gain scores, with standard errors, are also shown for various cohorts. "No math" indicates the cohort of students without college-level calculus.

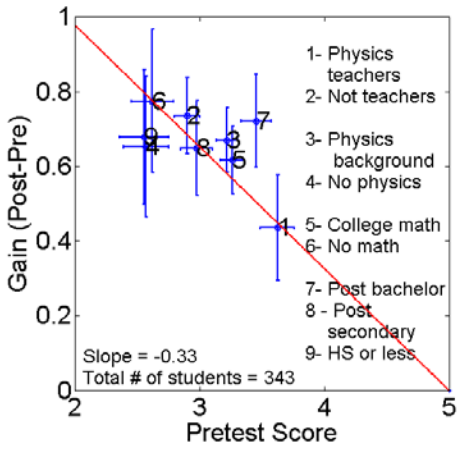

Figure 2. The negative slope of the red line, constrained to go through the point (5,0), indicates the normalized gain that best fits the 343 students who answered all 5 non-force-related items on both pre and post tests. The mean pretest and gain scores, with standard errors, are also shown for various cohorts.

Thus we have observed learning as measured by normalized gain that is between these limits. While both of our gains, 0.30 and 0.33 (+/-0.02), are closer to the gains Hake reported for traditional on-campus courses, they lie above *all* of the 14 traditional classes studied by Hake, suggesting that our students learn conceptual topics slightly better than in a traditional, lecture-based, class. This comparison is blunted by the fact that typically 19% of the first responses to a question were *preceeded by reference* to in-course resources, about a 1:1 ratio with the percentage of wrong answers. (Previously we found this ratio to be 1:3 in spite of a penalty that served to discourage students from giving wrong answers [Lee, Palazzo, Warnakulasooriya, & Pritchard, 2008].) More investigation shows several differences between student behavior on pre and posttest. Good comparisons of MOOCs and traditional courses with pre- and posttests must await a MOOC testing platform that prevents students from visiting other sites or materials before making their first response.

It is noteworthy that constraining the fit to pass through zero for a perfect score on the pretest does not add significantly to the uncertainty and thus is consistent with the data. We have termed such a situation one of "pure learning" (Pritchard, Lee, & Bao, 2008), which means that the data are consistent with the hypothesis that the number of initially incorrect answers given by each student (or each cohort) is reduced by a fraction equal to the normalized gain. The fact that no cohorts lie significantly below or above the best fit line in Figure 2 and Figure 3 should allay concerns that less well prepared students cannot learn in MOOCs.

Our concept test spanned a wider range of topics than Hake's study, which concerned topics covered in the first four weeks of 8.MReV. Furthermore, the questions on the Force Concept Inventory are narrow in scope, unlike the non-force questions in our pre-

and posttests that required analyzing answers, providing reasons for conceptual judgments, and finding the mistake in a given solution. However, the normalized gain seems insensitive to the particular questions used. Indeed, Hake's study reached its conclusions by amalgamating results from both the Force Concept Inventory and the Mechanics Baseline Test (which had some non-conceptual problems requiring choice among numerical answers and some questions on topics beyond the FCI). This insensitivity is further emphasized by the similarity of our results for force questions and all other questions (see Figures 1 and 2).

Relative Performance of Various Cohorts

The large enrollment and diverse demographics of the MOOC student body allow us to separately analyze and compare the relative learning of various cohorts of students in 8.MReV. We have formed and analyzed cohorts according to highest degree attained, academic preparation, and average ability in the course. Specifically, background in introductory physics and first semester calculus were used for academic preparation.

We have found one group that has a significantly higher normalized gain than all the rest. It is those 176 students who answered the two questions (12 and 13) on the pre and posttests that required a symbolic answer. Although their gains on those two questions were both less than 0.3, their gains on the other five non-force questions were sufficiently high that, for the seven questions, they had a normalized gain of 0.41 (+/- 0.03), as shown in Figure 3. The group has members from all previously mentioned cohorts in it, and is not distinguished by any obvious demographics.

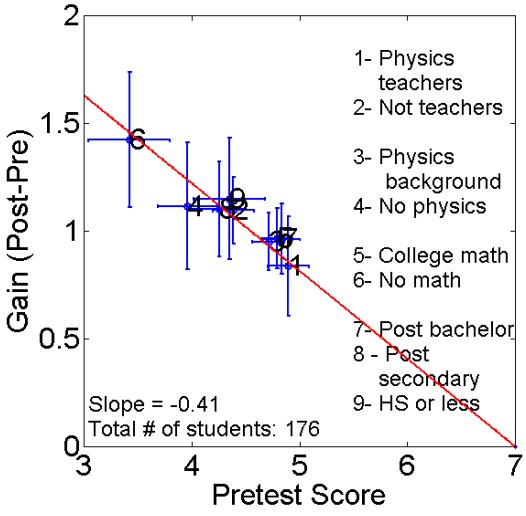

Figure 3. The negative slope of the red line, constrained to go through the point (7,0), indicates the normalized gain that best fits the 176 students who answered five non-force-related and two symbolic items on both pre and posttests. The mean pretest and gain scores, with standard errors, are also shown for various cohorts.

Figures 1-3, showing the normalized gain of the students in 8.MReV, display the pretest score and learning gains of the various cohorts just listed. Whether we look at gains on force-related concepts, non-force-related concepts, or examine the subset of each cohort that attempted the questions requiring the use of the symbolic answer interface in edX, we see no cohorts lying significantly below or above the normalized gain lines that fits all students in that sample. This certainly should allay concerns that less well prepared students cannot learn in MOOCs. In fact, the actual score improvement (gain) is *higher* for students with lower scores on the pretest (see Figures 1-3).

Pre- and Posttest Analysis with IRT

The pre- and posttest items occurred twice in the course, but the calibration process (to find their difficulty and discrimination) assumed they occur only once. For calibration we only used the responses to the posttest items, not the pretest, because the knowledge of the students at pretest would reflect details of the obviously highly variable previous instruction of our students. The resulting IRT discrimination and difficulty estimates were then used to calculate initial ability estimates for the students based on the pretest items.

Unlike the normalized gain procedures, because IRT can compare students who have taken different sets of items, we did not have to exclude as many students for the comparison, as we did with the normalized gain analysis. However, to obtain a reasonable estimate of a student's IRT score on the pretest and posttest, we only included the 578 students who had responded to at least seven pre- and eight posttest items. The gain in IRT score from the pretest to the posttest was 0.41 (standard error = 0.03) and independent of the average ability of the students, as seen in Figure 4.

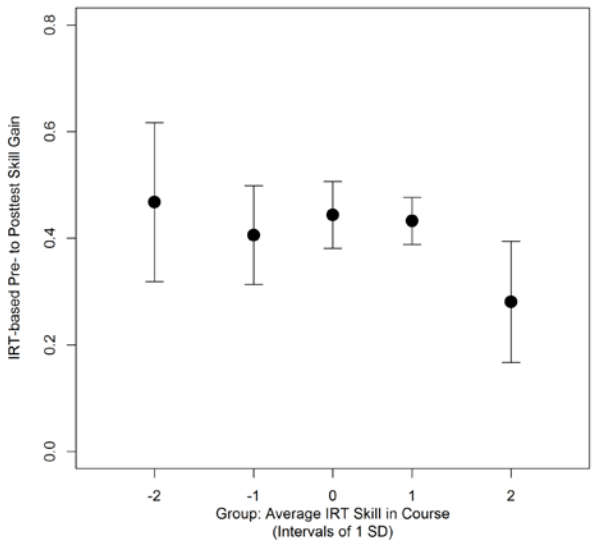

Figure 4. The IRT-based pre- to posttest gain for students grouped by their overall ability (skill) in the course.

Comparing MOOC and On-Campus Class Using Online Homework

Students in 8.MReV answer more homework questions each week than on the pre- or posttest. These responses therefore permit us to find the weekly IRT abilities of the students with reasonable statistical error. These give us an ability to quantify the improvement (or worsening) of the ability of students and cohorts over the semester. Ideally this will lead to insights about the learning habits and resource usage of students who improve relative to those who do not. This also allows us to measure the on-campus students in 8.011 (spring of 2013) on the scale of the 8.MReV MOOC students by using the online homework done by both groups. Of the 403 items, in the 8.MReV MOOC, 253 items had previously been given to the on-campus 8.011 students via the edX platform. The details of homework administration for both groups was highly similar: Both groups were allowed multiple attempts, both used the same platform, and both were done in an open-book, open Internet environment. Before we compare the overall ability and the week-by-week evolution of the abilities of the students in various cohorts of 8.MReV, including the 8.011 students as an additional cohort, we briefly describe the 8.011 course.

Description of On-Campus Course: 8.011

The on-campus course, 8.011, is the spring version of Introductory Newtonian Mechanics at MIT. This course, together with the subsequent Electricity and Magnetism course, is required of all MIT graduates, and most take it in their first semester. Students who earn less than a C in the fall Mechanics course are required to retake the course before moving on to Electricity and Magnetism; these students make up about 80% of the population of 8.011. In spring 2013, there were 47 students in 8.011, the first time the online segment of the course was run on the edX platform rather than on LON-CAPA. Of these 47 students, 35 attempted more than half of the online problems. Data from these 35 students were used in this study.

The course is designed primarily to help students review their understanding of the various topics in mechanics, but especially to help them organize this knowledge under five core models, which subsequently enables them to decide which core models apply to a particular problem. The pedagogy used is modeling applied to problem solving (Pawl et al., 2009). The first nine weeks of the course review the core topics and concepts in mechanics, and the subsequent six weeks involve topics with problems that require using several physics laws at once.

Comparative Abilities and Weekly Evolution of Cohorts

Figure 5 shows the distribution of student abilities on the items common between the two courses. The top-performing cohort, physics teachers, is highlighted. The teachers scored about half of a standard deviation above average, with a very few in the low-ability tail. The on-campus 8.011 students' ability averaged about 1.0 standard deviation

below the average in 8.MReV. In retrospect, this may not be surprising as the average 8.MReV student is far better educated, older, and is not juggling three or four other MIT courses.

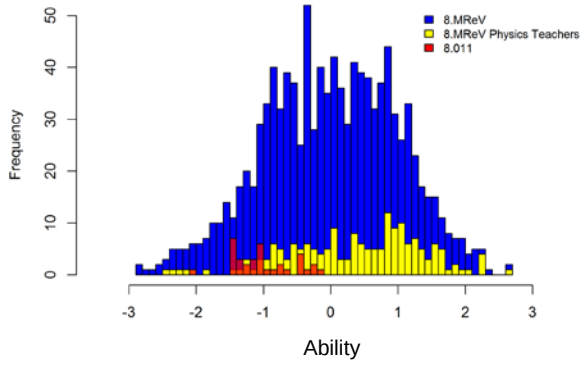

Figure 5. The distribution of abilities in 8.MReV overall, the teacher cohort in 8.MReV (yellow), and the on-campus 8.011 students (red).

Table 1 compares the relative abilities of various cohorts of students in 8.MReV with each other and with the on-campus students. We include cohorts based on their levels of education and status as 8.011 on-campus students or physics teachers. The relative improvement is the difference in a student's beginning and ending ability in the class as defined by a line of best to the student's weekly ability based on homework and quiz items. The relative improvement of a cohort is the average of the individual improvements.

Table 1

Average Ability of Various Cohorts

Cohort	Average ability in 8.MReV				
	% of 8.MReV	Mean	Standard deviation	Standard error	Relative improvement (SE)
PhD	8	0.67	0.93	.10	0.16 (0.06)
Masters	19	0.26	0.91	.06	-0.06 (0.05)
College	29	-0.08	0.99	.06	-0.11 (0.04)
High school	11	-0.20	0.93	.07	-0.11 (0.06)
Less than HS	6	-0.05	0.84	.10	-0.21 (0.10)
No response	23	0.02	1.04	.07	0.01 (0.07)
Physics teachers	17	0.39	0.97	.07	0.00 (0.05)
8.011 students	3	-1.05	0.50	.08	-

Note. Degree listed is highest degree attained. For example, "High School" refers to students who have obtained a high school diploma and may be enrolled in college.

The most salient feature of Table 1 is that the magnitude of the relative improvement, or change in relative ability, is small for all cohorts. A change of 0.2 would be less than the change from a B to a B+, for example. There is a small positive correlation between relative improvement and increased educational attainment. This might reflect that obtaining more education either develops or selects for greater learning skills. On the other hand, students with more education may, on average, have forgotten more of the physics they learned at a younger age so that they were able to relearn the material faster than younger students were able to learn it.

The graph in Figure 6 shows the weekly IRT abilities of several cohorts of 8.MReV students compared with the 8.011 on-campus students. Each set of weekly abilities was forced to have a mean of 0 and a standard deviation of 1. Therefore, maintaining the same ability across the course is not an indication of "not learning," but rather of not changing your ability relative to the class average. None of the cohorts shown in Figure 6 had a statistically significant relative improvement over the course. In particular, there was no evidence that cohorts with low initial ability learned less than the other cohorts.

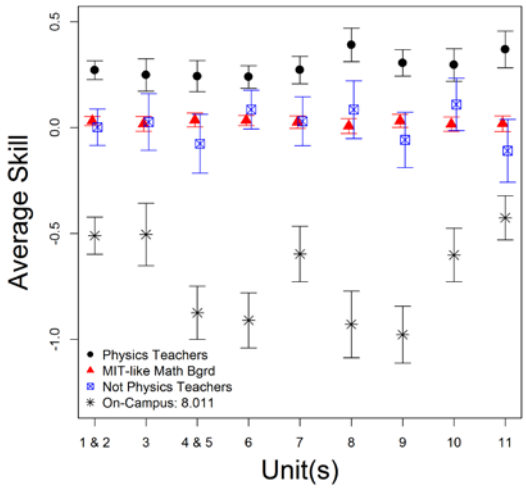

Figure 6. Weekly performance of various cohorts in 8.MReV compared with on-campus students in 8.011. Note: MIT-like Math Background refers to students with at least one-semester of calculus.

Comparison of On-Campus and MOOC Relative Ability

First, we address the question of whether the on-campus students measurably benefit from this environment. On-campus students, unlike the MOOC participants, benefitted from the following: four hours of instruction in which staff interacted with small groups of students (a flipped classroom) each week, staff office hours, helpful fellow students, available physics tutors, and the MIT library. Although the online students had lively discussions on many of the discussion boards that followed each eText page and problem and were required to do about 30% more problems (including more problems in the online tests than the on-campus students did in their weekly in-class tests), we still thought it likely that the on-campus students would show an increase in week-by-week ability relative to the online students. Clearly, this is not the case. The fact that most of the on-campus students had started the fall semester mechanics course (8.01) in 2012, but dropped out or completed it with a failing grade may have given them extra training on the material in the first several units. Note also that the on-campus course extended four weeks beyond the end of the material covered in 8.MReV and included topics like the harmonic oscillator, planetary orbits, a review of important procedures, and a general review of all material. But the bottom line remains: In spite of the extra instruction that the on-campus students had, Figure 6 shows no evidence of positive, weekly relative improvement of our on-campus students compared with our online students.

Summary and Conclusions

We have studied conceptual learning in a MOOC by analyzing the results of pre- and posttesting in two ways: normalized gain and item response theory (IRT). Both methods show unequivocal evidence of learning. The amount of learning, normalized gain, 0.31 +/- 0.02, was higher than in any of the 14 traditional (i.e., lecture-based) courses studied by Hake 1998, but was in the lowest decile of courses whose classes included "interactive engagement" activities.

The diversity and large sample size of the MOOC enabled us to separate the learning of various cohorts. We divided the sample into cohorts based on educational level, amount of preparation in mathematics, in physics, and by whether they were teachers. None of our nine cohorts had normalized gain that differed significantly from 0.31. This applied to questions involving both force and motion (the most frequently measured concepts in other studies) and all other topics taken together.

An advantage of comparisons of cohorts within our MOOC is that concerns about the MOOCs selecting only highly motivated students, or about the special nature of pre- and posttesting in our online environment, apply equally to all of the students in the MOOC. Thus, comparisons of various cohorts within the same MOOC give reasonably definitive results.

In addition, we have compared the weekly IRT abilities of students in a reformed on-campus course incorporating a flipped classroom relative to those of the 8.MReV class and several cohorts of its students. There is certainly no evidence that the on-campus students' four hours of intimate contact with teaching staff increased their relative ability over the term.

Now that we can measure the learning in our MOOC, we are in a position to study what correlates with it. Indeed we have made a preliminary investigation (Champaign et al., 2014) finding significant positive correlations with time spent on several different resources, but with little differentiation between them. A second factor that might affect learning is study patterns; for example, we found dramatically different patterns of resource use when students did homework versus exams (Seaton, Bergner, Chuang, Mitros, & Pritchard, 2014). This raises the question of whether students following these (or other) patterns will show more or less learning.

Another interesting area for future research is comparisons with other introductory physics MOOCs that have done pre-post testing to measure conceptual learning. Preliminary results from the Georgia Institute of Technology MOOC, which emphasizes students analyzing videos, show significantly less gain than we find here (G. Schatz, personal communication, June 19, 2014), whereas a video-based MOOC at the University of Colorado, which emphasizes conceptual learning, has significantly more gain than we see here, at least on force and motion (M. Dubson, personal communication, June 20, 2014). Given the different demographics of those registering,

the different objectives of each course, and the significantly smaller percentage of certificate earners in these other MOOCs, direct comparisons will be challenging.

In the future we can sharpen the results of pre- and posttesting by administering the same questions to on-campus students and those in the MOOC, constrained by the concern that too much pretesting in the MOOC may make some students withdraw because they feel like guinea-pigs.

It is also important to note the many gross differences between 8.MReV and on-campus education. Our self-selected online students are interested in learning, considerably older, and generally have many more years of college education than the on-campus freshmen with whom they have been compared. The on-campus students are taking a required course that most have failed to pass in a previous attempt. Moreover, there are more dropouts in the online course (but over 50% of students making a serious attempt at the second weekly test received certificates) and these dropouts may well be students learning less than those who remained. The pre- and posttest analysis is further blurred by the fact that the MOOC students could consult resources before answering, and, in fact, did consult within course resources significantly more during the posttest than in the pretest.

In summary, our MOOC produced significant and roughly equal *learning* for all of the cohorts differentiated along several axes that strongly influence their overall *ability*:

- students with high school or less education versus those with advanced college degrees;
- students lacking good preparation in math and physics – both obviously important for success in this course – versus those with good preparation; and
- students who display low ability versus high ability on the pretest.

In addition, we find a small improvement, relative to the overall class, for cohorts with a more formal education.

Acknowledgments

We acknowledge support from a Google Faculty Award, MIT, and NSF. We thank Yoav Bergner and Fiona Hollands for their suggestions and thoughtful comments on this manuscript. We thank Yoav Bergner and Daniel Seaton for work on analysis software used for these data.

References

Attali, Y. (2010). Immediate feedback and opportunity to revise answers: Application of a Graded Response IRT Model. *Applied Psychological Measurement*, 35(6), 472–479. doi:10.1177/0146621610381755

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Fredericks, C., Rayyan, S., Teodorescu, R., Balint, T., Seaton, D., & Pritchard, D. E. P. (2013). *From flipped to open instruction: The Mechanics Online Course*. A paper presented at the Sixth International Conference of MIT's Learning International Networks Consortium.

Hake, R. R. (1998). Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64-74.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hestenes, D. & Wells, M. (1992). A mechanics baseline test. *The Physics Teacher*, *30*, 159-165.

Hestenes, D. Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*, 141-158.

Hollands, F. M. & Tirthali, D. (May, 2014). *MOOCs: Expectations and reality*. *Full report*. Center for Benefit-Cost Studies of Education, Teachers College Columbia University. Retrieved from http://cbcse.org/wordpress/wp-content/uploads/2014/05/MOOCs_Expectations_and_Reality.pdf

Lee, Y.-J., Palazzo, D. J., Warnakulasooriya, R., & Pritchard, D. E. (2008). Measuring student learning with item response theory. *Physical Revew Special Topics: Physics Education Research*, *4*, 010102, DOI: 10.1103/PhysRevSTPER.4.010102

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics: Development, analysis, and application of psychological and educational tests*. The Hague, Netherlands: Eleven International.

Meyer, J. P., & Zhu, S. (2013). Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating. *Research and Practice in Assessment*, *8*, 26-39.

Pappano, L. (2012, November 2). The year of the MOOC. *The New York Times*.

Pawl, A., Barrantes, A., & Pritchard, D. (2009). Modeling applied to problem solving. *Proceedings of the 2009 Physics Education Research Conference*, 51-54.

Pawl, A., Barrantes, A., Cardamone, C., Rayyan, S. & Pritchard, D. E. (2011). Development of a mechanics reasoning inventory. *Proceedings of the 2011 Physics Education Research Conference*, *1413*, 287-290.

Pritchard, D. E., Lee, Y. J., & Bao, L. (2008). Mathematical learning models that depend on prior knowledge and instructional strategies. *Physical Review Special Topics-Physics Education Research*, *4*(1), 010109.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Seaton, D. T., Bergern, Y, Chuang, I, Mitros, P, & Pritchard, D. E. (2014). Who does what in a massive open online course? *Communications of the ACM*, *57*(4), 58-65.

Teodorescu, R., Pawl, A., Rayyan, S., Barrantes, A., & Pritchard, D. (2010). Toward an integrated online learning environment. *Proceedings of the 2010 Physics Education Research Conference*, *1289*, 321-324.

Thissen, D. (1991). MULTILOG: Multiple category item analysis and test scoring using item response theory [Computer software]. Chicago: Scientific Software International.

Appendix

Table 2

Pre- and Posttest Problems

	Problem summary	% Correct Pre	% Correct Post	Normalized gain	Attempts allowed
1	identify correct free body diagram, forces labeled	94	93	-0.13	4
2	identify correct free body diagram, forces not labeled	50	55	0.10	4
3	identify forces in free body diagram	63	68	0.15	4
4	answer sense-making: block on ramp attached to massive pulley	57	84	0.63	2
5	inelastic collision, find the error in given solution	46	56	0.19	1
6	explain answer to question above	62	74	0.33	2
7	pendulum swings down, collides inelastically. decomposition	75	67	-0.30	1
8	justify answer to question above	63	64	0.03	2
9	find force of rope pulling elevator at constant speed	76	85	0.40	2
10	direction of acceleration for block at bottom of circular ramp	44	68	0.43	3
11	direction of acceleration after block leaves ramp	84	91	0.43	3
12	find maximum elongation of mass on vertical spring after sudden release	58	65	0.16	3
13	heat generated after above mass comes to rest	55	61	0.13	3
14	find scale reading for mass in elevator given elevator's change in velocity over time interval	NA	63	NA	2
15	find maximum speed of cylinder on turntable, given m, mu, r	NA	81	NA	2
16	find maximum stretch of spring given initial position and velocity of given mass	48	68	0.38	10
17	moving mass approaches fixed mass with mutual attraction; find position where they collide given time of collision (use dynamics of center of mass)	27	44	0.23	10

Note. Percent correct based only on those students who attempted the problem.

The first 11 questions on the pre- and posttests are all multiple choice questions, generally involving conceptual issues rather than computations. These may be considered as conceptual tests, like the pioneering Force Concept Inventory (FCI) (Hake et al., 1992). The physics education research community has long compared the effectiveness of different pedagogies by the normalized gain on concept tests, that is, the fractional reduction in the number of incorrect responses to the questions on the posttest relative to the number of incorrect responses on the pretest.

Athabasca University