Designing Adaptive Assessments in MOOCs

Yigal Rosen
Harvard University
Cambridge, USA
yigal_rosen@harvard.edu

Ilia Rushkin
Harvard University
Cambridge, USA
ilia_rushkin@harvard.edu

Andrew Ang
Harvard University
Cambridge, USA
andrew_ang@harvard.edu

Colin Federicks
Harvard University
Cambridge, USA
colin_federicks@harvard.edu

Dustin Tingley
Harvard University
Cambridge, USA
dtingley@gov.harvard.edu

Mary Jean Blink
TutorGen, Inc.
Fort Thomas, USA
mjblink@tutorgen.com

ABSTRACT

There is an indisputable need for evidence-based instructional designs that create the optimal conditions for learners with different knowledge, skills and motivations to succeed in MOOCs. The study explores the technological feasibility and implications of adaptive functionality to course (re)design in the edX platform. Additionally, the study aims to establish the foundation for future study of adaptive functionality in MOOCs on learning outcomes, engagement and course drop-out rates. Preliminary findings suggest that the adaptivity of this kind leads to a higher efficiency of learning: students go through the course faster and attempt fewer problems, since the problems are served to them in a targeted way. And yet there is no evidence that the students' overall performance in the course suffers. Further research is needed to explore additional facets of adaptive assessment in different contexts of MOOCs and the effects on learning outcomes.

Author Keywords

MOOCs; assessment; adaptive assessment; adaptive learning.

INTRODUCTION

Digital learning systems are considered adaptive when they can dynamically change to enhance learning in response to student interactions within the MOOC rather than on the basis of preexisting information such as a learner's gender, age, or achievement test score. Adaptive learning systems use information gained as the learner works with them to vary such features as the way a concept is represented, its difficulty, the sequencing of problems or tasks, and the nature of hints and feedback provided. Adaptive technologies build on decades of research in intelligent tutoring systems, psychometrics, cognitive learning theory and data science [1, 3, 4]. These capabilities result in the ability to pinpoint the optimal pieces of content for learners (e.g., video, reading, discussion post, assessment item) across all educational domains based on growing evidence from the learner's performance and associated learning progression (i.e, learning objectives map). Harvard University partnered with TutorGen to explore the feasibility of adaptive learning and assessment technology implications of adaptive functionality to course (re)design in HarvardX, and examine the effects on learning outcomes, engagement and course drop-out rates. As the collaboration evolved, the following two strategic decisions have been made: (1) Adaptivity will be limited to assessments in four out of 16 graded sub-sections of the course. Extra problems will be developed to allow adaptive paths; and (2) Development efforts will be focused on Harvard-developed Learning Tool Interoperability (LTI) tool to support assessment adaptivity on edX platform. Therefore, in the current prototype phase of this project, adaptive functionality is limited to altering the sequence of problems. The order is determined by a personalized learning progression , using learners' real-time performance and statistical inferences on sub-topics they have mastered. The inferences are continuously updated based on each learners' performance.

While the prototype will enable us to explore the feasibility of adaptive assessment technology and implications of adaptive functionality to course (re)design in HarvardX, it will be challenging to anticipate its effects on learning outcomes, engagement and course drop-out rates due to the prototype limitations. However, we believe that the study will help to establish a solid foundation for future research on the effects of adaptive learning and assessment on outcomes such as, learning gains and engagement.

METHOD

A number of subsections in the course contain homework assessment pages, each made of several problems. The course users were randomly split 50%-50% into an experimental group and into a control group. When arriving on a homework page, users in the control group see a

L@S 2017, April 20 - 21, 2017, Cambridge, MA, USA

ACM 978-1-4503-4450-0/17/04…$15.00
DOI: http://dx.doi.org/10.1145/3051457.3053993

predetermined, non-adaptive set of problems on a page. In the experimental group, the experience is the same in all homework assessments except the four used in this study, where the adaptive tool was deployed. In those four assessments a user from the experimental group is served problems sequentially, one by one, in the order that is determined on-the-fly based on the user's prior performance. To enable adaptivity, all problems in the course were manually tagged with one or several learning objectives. Moreover, all problems in the 4 adaptive assessments were tagged with one of three difficulty levels: advanced, regular and easy. The adaptive engine (a variety of Bayesian Knowledge Tracing algorithm) decides which problem to serve next based on the list of learning objectives covered by the homework and course material. It estimates the user's mastery of a learning objective each time the user gives an answer to a problem tagged with this learning objective (even if this problem is outside of the adaptive assessments). If the problem served is advanced, the engine serves the instructional system advanced materials covering the necessary learning objectives, providing the students with an option to study these before attempting the problem. A given user in the experimental group does not necessarily see all of these problems. The user may stop working on the homework after reaching the required score (higher score does not give extra credit), or indeed for any other reason. In addition, the engine may stop serving problems if the user's mastery level for a learning objective becomes sufficiently high that it needs no further verification. Students in the control group also have access to these materials in an optional part of the course.

In order to explore possible effects of adaptive experiences on learners' mastery of content knowledge competence-based pre- and post-assessment were added to the course and administered to study participants in both experimental and control groups. Typical HarvardX course clickstream time-stamped data and pre-post course surveys data will also be collected and analyzed.

Course Design Considerations

Adaptive learning techniques require the development of additional course materials, so that different students can be provided with different content. For our prototype, tripling the existing content in the four adaptive subsections was considered a minimum to provide a genuine adaptive experience. This was achieved by work from the project lead and by hiring an outside content expert. The total time outlay was ~200 hours. Keeping the problems housed within the edX platform avoided substantial amounts of software development.

LTI Tool Development

To enable the use of an adaptive engine in an edX course, Harvard developed the Bridge for Adaptivity (BFA) tool. BFA is a web application that uses the LTI specification to integrate with learning management systems such as edX. BFA acts as the interface between the edX course platform and the TutorGen SCALE (Student Centered Adaptive Learning Engine) system, and handles the display of problems recommended by the adaptive engine.

This LTI functionality allows BFA to be embedded in one or more locations in the course. The user interface seen by a learner when they encounter an installed tool instance is shown below:

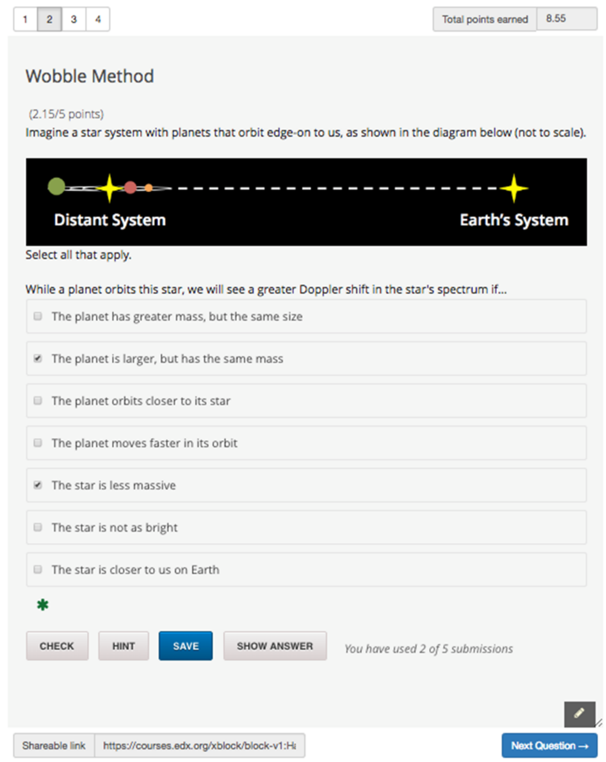

Figure 1. Adaptive assessment user interface

Problems from the edX course are displayed one at a time in a center activity window, with a surrounding toolbar that provides features such as navigation, a score display, and a shareable link for the current problem (that the learner can use to post to a forum for help). When a learner completes a problem in the activity window, embedded Javascript in the edX content sends data about the learner and their response to BFA. This data is then processed and sent to SCALE. When the learner chooses to advance to the next problem, BFA makes a query in real-time to SCALE for the next recommended activity for that learner, then serves the appropriate edX content in the activity window via xBlock URL.

TutorGen Adaptive Engine

TutorGen SCALE, is focused on improving learning outcomes using data collected from existing and emerging educational technology systems combined with the core technology to automatically generate adaptive capabilities. Key features that SCALE provides include knowledge tracing, skill modeling, student modeling, adaptive problem selection, and automated hint generation for multi-step problems. SCALE engine it improves over time with additional data and/or with the help of human input by providing machine learning using a human centered approach The algorithms have been tested on various data sets in a wide range of domains. For successful implementation and optimized adaptive operations, it is important that the knowledge components / skills (KC) be tagged at the right level of granularity. The system will provide opportunity to refine the tagging of these KCs after data has been collected from actual student interactions.

SCALE has been used in the intelligent tutoring system environment, providing adaptive capabilities during the formative learning stages. SCALE with HarvardX for this course is being used more as in the assessment stage of the the student experience. In order to accomplish the goals of the prototype for this pilot study, we extended our algorithms to consider not only the learning objectives, identified as the KCs, but also to consider problem difficulty and problem selection within the modules or groupings of concepts and problems. This will accommodate the needs for this course by providing an adaptive experience for students while still supporting the logical flow of the course. Further, the flexible nature of the course, having all content available and open to students for the duration of the course, presents some additional requirements to ensure that students are presented with problems based on their current state and not necessarily where the system believes they should navigate.

PRELIMINARY FINDINGS

The course was launched on Oct 19, 2016. The data for the analysis presented in this paper were accessed on Jan 04, 2017 (plus or minus a few days, since different parts of the data were extracted at different times), about two and a half months later.

	Experimental group	Control group
Regular level only	68	99
Easy level only	0	0
Advanced level only	1	1
(Regular ∪ Easy) levels only	2	41
(Regular ∪ Advanced) levels only	121	0
(Easy ∪ Advanced) levels only	0	1
(Regular ∪ Easy ∪ Advanced) levels	84	144
Total students attempting new problems	276	286

Table 1. Number of students attempting assessment items of different difficulty level

More students are registering for the course on a daily basis, so the results of the analysis are preliminary. We will refer to the list of problems from which problems were served adaptively to the experimental group as "new problems". The control group may have interacted with these as well, although not adaptively. There were 39 new problems, out of which 13 were regular difficulty (these formed the assessments for the control group of students), 14 were advanced and 12 were easy. For the control group, the advanced and easy problems were offered as extra material after assessment, with no credit toward the course grade. The numbers of students attempting assessment problems of different difficulty levels are given in Table 1.

To get a sense of how the two groups of students performed in the course, we compared the group averages of the differences in scores in the pre-test and post-test (Figure 2). We included only the scores from the test questions tagged with the learning objectives that are encountered among the new problems. Each question was graded on the scale 0-1, and we took the average question score for each student in each test.

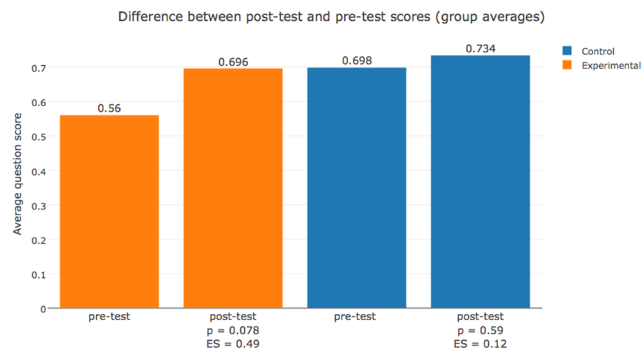

Figure 2. Comparison of post-test and pre-test scores. The population of users is subset to only those who attempted the pre-test, the new problems, and the post-test. Here and everywhere below, the p-values are two-tailed from the Welch two-sample t-test, and the effect size is the Cohen's d.

There is a noticeable between-group difference in the pre-test scores (p-value 0.066, effect size 0.46). This is due to subsetting to those users who attempted new problems and the post-test (in the absence of this subsetting, the effect size drops to 0.00028, meaning that initially the two populations have virtually no difference, as expected), Therefore, Figure 2 shows two patterns: 1) the experimental group achieves a larger knowledge gain, even with less prior knowledge; 2) in the experimental group students with low prior knowledge are more likely not to drop out and reach the post-test.

We did not see a difference in the final grade of the course: the mean grade was 84.3% in the experimental group vs. 85.77% in the control group, which is not significant at all (p-value 0.63, effect size −0.12).

Students in the experimental group tended to make more attempts at a problem (Figure 3). they tried fewer problems (Figure 4) most strikingly among the easy new problems: for these we have 1,122 recorded scores in the control group and only 325 in the experimental group. The interpretation emerges that the students who experienced adaptivity showed more persistence by giving more attempts per problem (presumably, because adaptively served problems are more likely to be on the appropriate current mastery level for a student), while taking a faster track through the course materials. Corroborating this last interpretation, we observe that the experimental group students tended to have a lower net time on task in the course: an average of 4.37 hours vs. 4.80 in the control group (in this comparison, p-value 0.11, effect size –0.14).

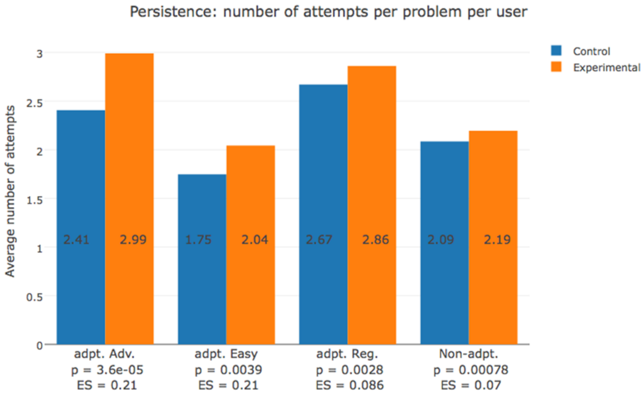

Figure 3. Comparison of attempt numbers between the experimental and control groups in the modules (chapters) where adaptivity was implemented. The attempt numbers are averaged both over the problems and over the users.

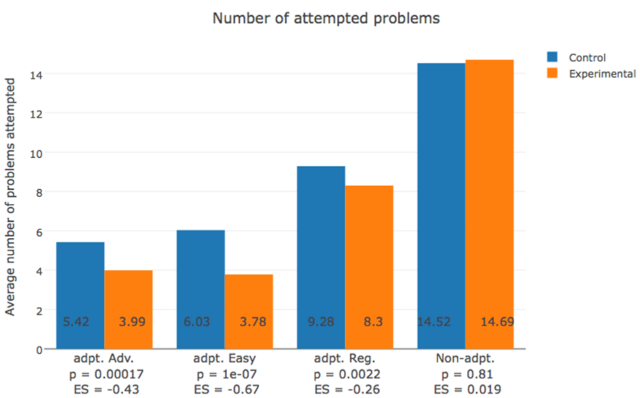

Figure 4. Comparison of attempt numbers between the experimental and control groups in the modules (chapters) where adaptivity was implemented.

No significant between-group difference was found in the rates of course completion and certification, or in demographics of students who did not drop out.

Thus, we propose that the adaptivity of this kind leads to a higher efficiency of learning: students go through the course faster and attempt fewer problems, since the problems are served to them in a targeted way. And yet there is no evidence that the students' overall performance in the course suffers: in fact, Figure 2 tentatively suggests a benefit. Given the limited implementation of adaptivity in this course, it is not surprising that we cannot find a statistically significant effect on student overall performance in the course. We expect to refine these conclusions in the future courses with a greater scope of adaptivity.

FUTURE WORK

There appear to be extensive opportunities to expand adaptive learning and assessment in MOOCs. Ideally, larger sets of questions that are tagged to the learning objectives for a module could provide a more adaptive learning experience for students, while also providing a higher degree of certainty of assessment results. Given the structure of many MOOCs, more integration between learning content and assessment could provide an adaptive experience that would guide students to content that could improve their understanding based on how they perform on integrated assessments. Affective factors, such as boredom and frustration, as well as behaviors like gaming the system, are areas where, if detected, the system could provide a more personalized learning experience. Finally, this work could lead to improved MOOC platform features that would contribute to improved student experiences, such as optimized group selection [2]. In addition, we anticipate expanding this adaptive assessment system to work with other LTI-compliant course platforms. Enabling use in a platform such as Canvas, the learning management system used university-wide at Harvard (and many other schools), would enable adaptivity for residential courses on a large scale. An adjustment to the current system architecture would be the use of OpenEdX as the platform for creating and hosting problems.

ACKNOWLEDGEMENTS

We are grateful for the support from the Office of the Vice Provost for Advances in Learning at Harvard University for thoughtful leadership and support to HarvardX and the VPAL-Research group. Special thanks to Professor Dimitar Sasselov from the Department of Astronomy whose SuperEarths and Life MOOC make this project possible.

REFERENCES

1. Koedinger, K., and Stamper, J. 2010. A Data Driven Approach to the Discovery of Better Cognitive Models. In Baker, R.S.J.d., Merceron, A., Pavlik, P.I. Jr. (Eds.) *Proceedings of the 3rd International Conference on Educational Data Mining*. (EDM 2010), 325-326. Pittsburgh, PA.
2. Rosen, Y. in press. Assessing students in human-to-agent settings to inform collaborative problem-solving learning. *Journal of Educational Measurement*.
3. Rosen, Y. 2015. Computer-based assessment of collaborative problem solving: Exploring the feasibility of human-to-agent approach. *International Journal of Artificial Intelligence in Education* 25, 3: 98-129.
4. Stamper, J., Barnes, T., and Croy, M. 2011. Experimental Evaluation of Automatic Hint Generation for a Logic Tutor. In Kay, J., Bull, S. and Biswas, G. eds. *Proceeding of the 15th International Conference on Artificial Intelligence in Education (AIED2011)*. 345-352. Berlin Germany: Springer.