Adaptive Assessment Experiment in a HarvardX MOOC

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

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

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

Colin Fredericks
Harvard University
Cambridge, USA
colin_fredericks@harvard.edu

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

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

Glenn Lopez
Harvard University
Cambridge, USA
glenn_lopez@harvard.edu

ABSTRACT

We report an experimental implementation of adaptive learning functionality in a self-paced HarvardX MOOC (massive open online course). In MOOCs there is need for evidence-based instructional designs that create the optimal conditions for learners, who come to the course with widely differing prior knowledge, skills and motivations. But users in such a course are free to explore the course materials in any order they deem fit and may drop out any time, and this makes it hard to predict the practical challenges of implementing adaptivity, as well as its effect, without experimentation. This study explored the technological feasibility and implications of adaptive functionality to course (re)design in the edX platform. Additionally, it aimed to establish the foundation for future study of adaptive functionality in MOOCs on learning outcomes, engagement and drop-out rates. Our preliminary findings suggest that the adaptivity of the kind we used leads to a higher efficiency of learning (without an adverse effect on learning outcomes, learners go through the course faster and attempt fewer problems, since the problems are served to them in a targeted way). Further research is needed to confirm these findings and explore additional possible effects.

Keywords

MOOCs; assessment; adaptive assessment; adaptive learning.

1. INTRODUCTION

Digital learning systems are considered adaptive when they can dynamically change the presentation of content to any user based on the user's individual record of interactions, as opposed to simply sending users into different versions of the course based on preexisting information such as user's demographic information, education level, or a test score. Conceptually, an adaptive learning system is a combination of two parts: an algorithm to dynamically assess each user's current profile (the current state of knowledge, but potentially also affective factors, such as frustration level), and, based on this, a recommendation engine to decide what the user should see next. In this way, the system seeks to optimize individual user experience, based on each user's prior actions, but also based on the actions of other users (e.g. to identify the course items that many others have found most useful in similar circumstances). Adaptive technologies build on decades of research in intelligent tutoring systems, psychometrics, cognitive learning theory and data science [1, 3, 4].

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 were made: (1) Adaptivity would be limited to assessments in four out of 16 graded sub-sections of the course. Extra problems would be developed to allow adaptive paths; (2) Development efforts would be focused on Harvard-developed Learning Tools 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, based on continuously updated statistical inferences on knowledge components a user mastered. As a supplement to these assessment items, a number of additional learning materials are served adaptively as well, based on the rule that a user should see those before being served more advanced problems.

While the prototype enabled us to explore the feasibility of adaptive assessment technology and implications of adaptive functionality to course (re)design in HarvardX, it is still challenging to judge 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. [5]

2. SETUP AND USER EXPERIENCE

The HarvardX course in this experiment was “Super-Earths and Life”. It deals with searching for planets orbiting around stars other than the Sun, in particular the planets capable of supporting life. The subject matter is physics, astronomy and biology. Roughly speaking, the course aims at users with college-level knowledge of physics and biology. Some of the assessment material in the course requires calculations, and some requires extensive factual knowledge (e.g. questions about DNA structure).

Two versions of the course have already run in the edX platform, our adaptivity was implemented as part of the course re-design for the third run.

A number of subsections in the course contained assessment modules (homeworks). The experiment consisted of making four of these homeworks adaptive for some of the users. At the moment of their registration, the course users were randomly split 50%-50% into an experimental group and into a control group. When arriving to a homework, users in the control group see a predetermined, non-adaptive set of problems on a page. The same is true for the experimental group in all homeworks except the four where we deployed the adaptive tool. In these homeworks, a user from the experimental group is served problems sequentially, one by one, in the order that is individually determined on-the-fly based on the user's prior performance. In addition to problems, some instructional text pages were also included in the serving sequence.

To enable adaptivity, we manually compiled a list of knowledge components (KCs, for our purposes synonymous with "learning objectives", "learning outcomes", or "skills") and tagged problems in the course with one or several knowledge components. This tagging was done for *all* assessment items in the course (as well as for some learning materials), enabling the adaptive engine to gather information from any user's interaction with any problem in the course, not only with those problems that are served adaptively. Additionally, the problems in the 4 adaptive homeworks were tagged with one of three difficulty levels: advanced, regular and easy (other problems in the course were tagged by default as regular). No pre-requisite relationships or other connections among the knowledge components were used.

The adaptive engine (a variety of Bayesian Knowledge Tracing algorithm) decides which problem to serve next based on the list of KCs covered by the homework and course material. Additional rules could be incorporated into the serving strategy. Thus, we had a rule that before any problem of difficulty level "Advanced", the user should see a special page with advanced learning material.

The parity between experimental and control groups was set up as follows. In the pool from which problems are adaptively served to the experimental group, all the regular-difficulty problems were the ones that the control group saw in these homework. The control group had access to the easy and advanced problems as well: students in this group saw a special "extra materials" page after each of the 4 experimental homeworks. This page contained the links to all the advanced instructional materials and advanced and easy problems for this homework, for no extra credit. Thus, all the materials that an experimental user can see, were also available to the control students. There were two main reasons for this: obvious usefulness for comparative studies, and enabling all students, experimental and control, to discuss all problems in the course forum.

When an experimental group user is going through an adaptive homework, the LTI tool loads edX problem pages in an iFrame. Submitting ("checking") an answer to the problem triggers an update of user's mastery, but does not trigger serving the next problem. For that to happen, the user has to click the button "Next Question" outside the iFrame. The user always can revisit any of the previously served problems.

In edX, users usually get several attempts at a problem. Thus, it may be possible for a user to submit a problem after the next problem has already been served. Fig. 1, for instance, shows a situation, where so far 4 problems have been served (note the numbered tabs in the upper left), but the user is currently viewing problem 2 in this sequence, not the latest one. The user is free to re-submit this problem, which will update the user's mastery (although in this case there is no need to do so, since it appears that problem 2 has been answered correctly). It will not alter the existing sequence (problems 3 and 4 will not be replaced by others), but it may have effect on what will be served as 5 and so on.

The user interface keeps track of the total number of points earned in a homework (upper right corner in Fig. 1). The user knows how many points in total are required and may choose to stop once this is achieved (earning more points will no longer affect the grade). Otherwise, the serving sequence ends when the pool of questions is exhausted. Potentially, it could also end when the user's probability of mastery on all relevant KCs passes a certain mastery threshold (a high probability, at which we consider the mastery to be, in practical terms, certain; it was set to 0.9). However, in this particular implementation, due to having only a modest number of problems, this was not done.

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 was collected.

2.1 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. This did not provide each knowledge component with a large number of problems, reducing the significance of knowledge tracing, but it was sufficient for the purpose of our experiment. The total time outlay was ~200 hours. Keeping the problems housed within the edX platform avoided substantial amounts of software development.

The tagging of content with knowledge components was done by means of a shared Google spreadsheet, which contained a list of content items in one sheet (both assessment and learning materials), a list of knowledge components in another, and a correspondence table (the tagging itself), including the difficulty levels, in the third.

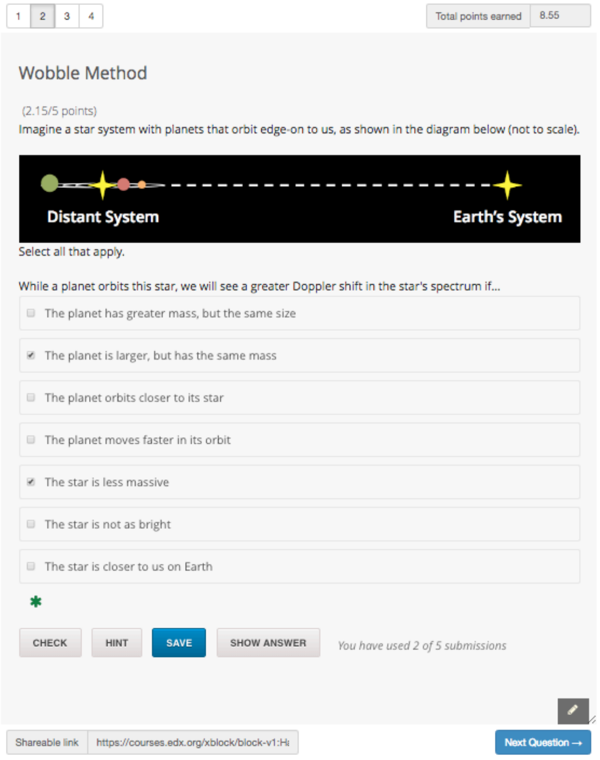
Most of the time was spent on creating new problems based on the existing ones. For these the tagging process was "reversed": rather than tag existing content with knowledge components, the experts created content targeting knowledge components and difficulty levels. Commonly, an existing problem was considered to be of "regular" difficulty, and the expert's task was to create an "easy" and/or an "advanced" version of it.

103 distinct knowledge components were used in tagging. The experts used their judgement in defining them. 66 of these were used in tagging problems, and in particular the 39 adaptively served problems were tagged with 25 KCs. The granularity of KCs was such that a typical assessment problem was tagged with one learning objective (which is desirable for knowledge tracing). Namely, among the adaptively served problems, 31 were tagged with a single KC, 7 problems – with 2 KCs, and 1 problem – with 3.

2.2 LTI Tool Development

To enable the use of an adaptive engine in an edX course, Harvard developed the Bridge for Adaptivity (BFA) tool (open-source, GitHub link available upon request). 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. Problems are accessed by edX XBlock URLs.

This LTI functionality allows BFA to be embedded in one or more locations in the course (4 locations in our case). The user interface seen by a learner when they encounter an installed tool instance is that shown in Fig. 1.



Wobble Method

(2.15/5 points)

Imagine a star system with planets that orbit edge-on to us, as shown in the diagram below (not to scale).

Select all that apply.

While a planet orbits this star, we will see a greater Doppler shift in the star's spectrum if...

- □ The planet has greater mass, but the same size
- ☑ The planet is larger, but has the same mass
- □ The planet orbits closer to its star
- □ The planet moves faster in its orbit
- ☑ The star is less massive
- □ The star is not as bright
- □ The star is closer to us on Earth

You have used 2 of 5 submissions

https://courses.edx.org/xblock/block-v1:H

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). The diagram in Fig. 2 describes the data passing in the system. The user-ids used by edX are considered sensitive information and are not shared with SCALE: we created a different user-id system for SCALE, and the mapping back and forth between the two id-systems happens in the back end of the app.

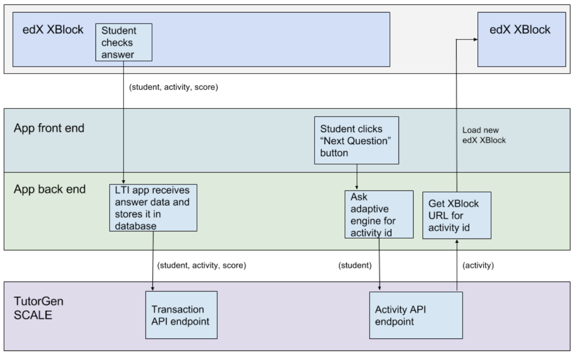

edX XBlock
Student checks answer
edX XBlock
(student, activity, score)
App front end
Student clicks "Next Question" button
Load new edX XBlock
App back end
LTI app receives answer data and stores it in database
Ask adaptive engine for activity id
Get XBlock URL for activity id
(student, activity, score)
(student)
(activity)
TutorGen SCALE
Transaction API endpoint
Activity API endpoint

Figure 2. Diagram of data passing in the system

Every problem-checking event by the user (both inside and outside the adaptive homeworks) sends the data to SCALE, to update the mastery information real-time. Every "Next Question" event in an adaptive homework sends to SCALE a request for the next content item to be served to the user (this could be instructional material or a problem). SCALE sends back the recommendation, which is accessed as an edX XBlock and loaded.

The edX support for LTI is highly stable. The challenge is that edX exports data on a weekly cycle, but we needed to receive the information about submits in real time. We achieved this by creating a reporting JavaScript and inserting it into every problem.

2.3 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 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 be tagged at the right level of granularity.

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 student experience. In order to accomplish the goals of the prototype for this pilot study, we extended our algorithms to consider not only the knowledge components (KCs), but also problem difficulty. 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.

A variety of serving strategies are available in SCALE and can be swapped in and out. In this particular implementation, while the algorithm did trace the students' knowledge, the results were used minimally in the serving strategy: it did not make sense to do otherwise given the small size of the adaptive problem pool. SCALE was configured to consider after each submit: the probability of the learner has mastered the KCs from the problem most recently worked, the difficulty of that problem, and the correctness of the submitted answer. A general and simplified explanation of the process is as follows. Each of the four adaptive modules was treated as a separate instance, with its own pool of problems. Each problem can be served to each learner no more than once. Given the last problem submitted by a learner in the module, the candidate to be served next is the (previously unseen) problem, whose KC tagging overlaps with the KCs of the last submitted problem and includes at least one KC, on which the user has not yet reached the mastery threshold. If multiple candidates are available, SCALE will serve the one with a KC closest to mastery. If no candidates are available, other problems of the same difficulty within the same module will be served (i.e. SCALE switches to another KCs). The difficulty level of the next served problem is determined by the last submit correctness. As long as problems of the same difficulty level as the last one are available, the learner will remain at that difficulty level. Once such problems are exhausted, SCALE will serve a more or less difficult problem, depending on whether the last submit in the module was correct or incorrect.

2.4 Quantitative Details and Findings

The course was launched on Oct 19, 2016. The data for the analysis presented in this paper were accessed on Mar 08, 2017 (plus or minus a few days, since different parts of the data were extracted at different times), after the official end date of the course.

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

	Experimental group	Control group
Regular level only	58	73
Easy level only	0	0
Advanced level only	1	0
(Regular ∪ Easy) levels only	1	35
(Regular ∪ Advanced) levels only	105	0
(Easy ∪ Advanced) levels only	0	1
(Regular ∪ Easy ∪ Advanced) levels	99	145
Total students attempting new problems	264	254

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 (as additional problems that do not count towards the grade). 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. For reasons unrelated to this study, both tests were randomized: in each test each user received 9 questions, randomly selected from a bank of 17. All questions were graded on the 0-1 scale. The users knew that the pre- and post- tests do not contribute to the grade, and so only about ~40% of users took both. Moreover, not all of these questions were *relevant* for (i.e. tagged with) those 25 knowledge components, with which the adaptively served problems were tagged. So the number of offered *relevant* questions varied randomly from user to user. For these reasons the pre- and post-test are not the most reliable measure of knowledge gain, but it was still important for us to make sure that adaptivity did not have any adverse effect. Each question was graded on the scale 0-1, and in Fig. 3 we subset the student population to those individuals who attempted a "new problem" *and* a relevant pre-test question *and* a relevant post-test question, and used the average score from relevant questions as the student's relevant score. For instance, if one user attempted two relevant questions in a pre-test, and another user attempted three, and the questions were answered correctly, both users have the relevant score 1: (1+1)/2=(1+1+1)/3.

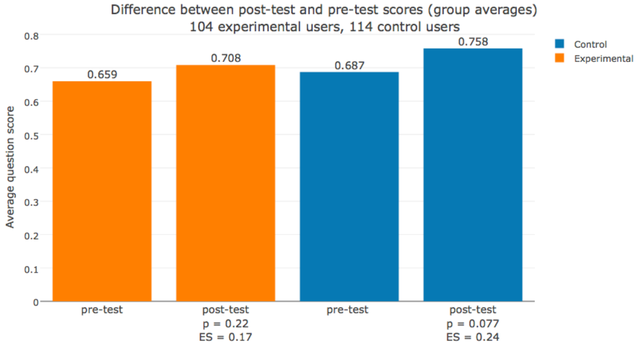


Figure 3. Comparison of relevant post-test and pre-test scores. Here and everywhere below, the p-values are two-tailed from the Welch two-sample t-test, and the effect size is the Cohen's d (Cohen suggested to consider d=0.2 as "small", d=0.5 as "medium" and d=0.8 as "large" effect size).

There is no significant between-group difference, neither in the pre-test scores (p-value 0.49, effect size 0.093) nor in the post-test scores (p-value 0.21, effect size 0.17). The two populations of pre-test takers remain comparable after subsetting to those who attempted new problems and the post-test and we see no statistically significant difference in the knowledge gaining between the experimental and control groups.

We did not see a difference in the final grade of the course: the mean grade was 83.7% in the experimental group vs. 82.9% in the control group, which is not a significant difference (p-value 0.76, effect size 0.06). Likewise, there is no significant between-group difference in the completion and certification rates (about 20%), or in demographics of students who did not drop out.

Students in the experimental group tended to make more attempts at a problem (Fig. 4), and they tried fewer problems (Fig. 5), most strikingly among the easy new problems: for these we have 1,162 recorded scores in the control group and only 423 in the experimental group.

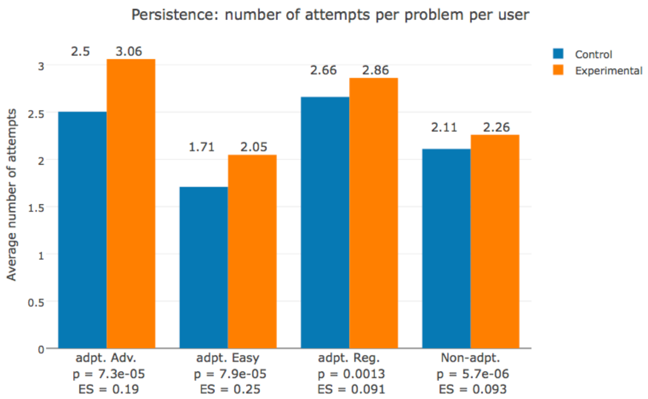


Figure 4. Comparison of attempt numbers between the experimental and control groups in the chapters where adaptivity was implemented. The attempt numbers are averaged both over the problems and over the users. Non-adaptive problems are problems not from the 4 experimental homeworks but from the same two chapters of the course as the experimental homeworks.

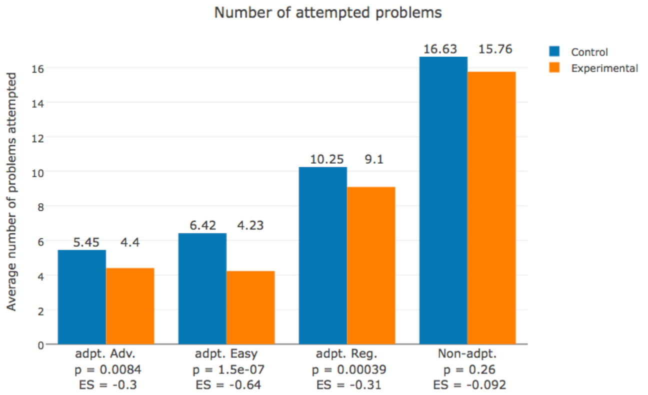


Figure 5. Comparison of attempt numbers between the experimental and control groups in the chapters where adaptivity was implemented. Non-adaptive problems are problems not from the 4 experimental homeworks but from the same two chapters of the course as the experimental homeworks.

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. We also observed that the experimental group students tended to have a lower net time on task in the course: an average of 5.47 hours vs. 5.85 in the control group (although in this comparison the p-value is high, 0.21, and the effect size is –0.11).

Thus, we conjecture 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 of an adverse effect on the students' overall performance or knowledge gain. 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.

3. FUTURE WORK

Our implementation of adaptivity provided some insights for future work. For instance, assessment questions in MOOCs can vary greatly in nature, difficulty and format (multiple choice, check-all-that-applies, numeric response, etc.), and may often be tagged with more than one knowledge component. To be suitable for a MOOC, an adaptive engine should be able to handle these features.

There appear to be extensive opportunities to expand adaptive learning and assessment in MOOCs. The low total number of problems was the most severe restriction on the variability of learner experience in this study. In the future applications, larger sets of tagged items could provide a more adaptive learning experience for students, while also providing a higher degree of certainty of assessment results. Interestingly, in some MOOCs (for example, those teaching programming languages) it may be possible to create very large numbers of questions algorithmically, essentially by filling question templates with different data.

In this study, adaptivity was implemented mostly on assessment problems. 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 could be included to provide a more personalized learning experience. We can conceive an adaptive engine which decides what item to serve next based not just on the mastery but also on the behavioral patterns interpreted as boredom or frustration.

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.

4. ACKNOWLEDGMENTS

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 Harvard Department of Astronomy whose "Super-Earths and Life" MOOC made this project possible. TutorGen gratefully acknowledges support of SCALE® from the National Science Foundation award numbers 1346448 and 1534780 and from the Commonwealth of Kentucky Cabinet for Economic Development, Kentucky Science and Engineering Foundation, and The Kentucky Science and Technology Corporation, award numbers KSTC-184-512-14-182 and KSTC-184-512-16-241.

5. 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. 2017. Assessing students in human-to-agent settings to inform collaborative problem-solving learning. *Journal of Educational Measurement* 54, 1: 36-53.

[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.

[5] A preliminary report of our study (based on the data obtained prior to the course end) is to appear in the *Proceedings of the Fourth Annual ACM Conference on Learning at Scale* (L@S 2017) as: Rosen, Y., Rushkin, I., Ang A., Fredericks C., Tingley D., Blink M.J. 2017. Designing Adaptive Assessments in MOOCs.