DOI:10.1145/2500876

Student-participation data from the inaugural MITx (now edX) course—6.002x: Circuits and Electronics—unpacks MOOC student behavior.

BY DANIEL T. SEATON, YOAV BERGNER, ISAAC CHUANG, PIOTR MITROS, AND DAVID E. PRITCHARD

Who Does What in a Massive Open Online Course?

MASSIVE OPEN ONLINE COURSES (MOOCs) collect valuable data on student learning behavior; essentially complete records of all student interactions in a self-contained learning environment, with the benefit of large sample sizes. Here, we offer an overview of how the 108,000 participants behaved in 6.002x - Circuits and Electronics, the first course in MITx (now edX) in the Spring 2012 semester. We divided participants into tranches based on the extent of their assessment activities, ranging from browsers (constituting ~76% of the participants but only 8% of the total time spent in the course) to certificate earners (7% of participants who accounted for 60% of total time). We examined how the certificate earners allocated their time among the various course components and what fraction of each they accessed. We analyze transitions between course components, showing how student behavior differs when solving homework vs. exam problems. This work lays the foundation for future studies of how various course components, and transitions among them, influence learning in MOOCs.

Though free online courses are not new,[8] they have reached an unprecedented scale since late 2011. Three organizations—Coursera, edX, and Udacity—have released MOOCs[13] drawing more than 100,000 registrants per course. Numbers from these three initiatives have since grown to more than 100 courses and three million total registrants, resulting in 2012 being dubbed "The Year of the MOOC" by the *New York Times*.[16] Though there has been much speculation regarding how these initiatives may reshape higher education,[6,12,20] little analysis has been published to date describing student behavior or learning in them.

Our main objective here is to show how the huge amount of data available in MOOCs offers a unique research opportunity, a means to study detailed student behavior in a self-contained learning environment throughout an

» key insights

- **Data collected in MOOCs provides insight into student behavior, from weekly e-textbook reading habits to context-dependent use of learning resources when solving problems.**
- **In 6.002x, 76% of participants were browsers who collectively accounted for only 8% of time spent in the course, whereas, the 7% of certificate-earning participants averaged 100 hours each and collectively accounted for 60% of total time.**
- **Students spent the most time per week interacting with lecture videos and homework, followed by discussion forums and online laboratories; however, interactions with the videos and lecture questions were distinctly bimodal, with half the certificate earners accessing less than half of these resources.**

ILLUSTRATION BY ANTHONY FREDA

entire course. We thus studied the approximately 100GB of time-stamped log data describing student interactions with the inaugural MITx course 6.002x Circuits and Electronics in spring 2012, data at least two orders of magnitude larger than was analyzed in previous studies of online learning.[10,21] We develop and exhibit several ways to study student interactions with course resources. We do not analyze demographic factors, but rather differentiate students by number of assessment items attempted and total time spent in the course. We studied all registrants with these metrics before turning to the more detailed time allocation and resource use of students earning a certificate of accomplishment. For certificate earners, we examined the use of course components (such as lecture videos, homework, and discussion forums) in terms of user time allocation and total fraction accessed. We also studied resource use during problem solving, revealing markedly different patterns of accesses and time allocation among different course components when students solve problems during homework vs. when taking exams.

The correlation of attrition with less time spent in early weeks begs the question of whether motivating students to invest more time would increase retention rates.

6.002x, Procedures, Data Analysis

With some modification for online delivery, the 14 weeklong units of 6.002x largely mirrored a traditional on-campus course in both format and timing. The course sequence (see Figure 1, left navigation bar) involves lecture sequences consisting of lecture videos (annotated PowerPoint slides and actual MIT lectures) with embedded lecture questions, tutorial videos (recitation substitute), homework (three to four multi-part problems), and lab assignments (interactive circuit toolbox). Overall grades were determined by homework (15%), labs (15%), a midterm (30%), and a final (40%). Supplementary materials (see Figure 1, top navigation bar) included a course textbook (navigable page images), a staff- and student-editable wiki and moderated student discussions. For further exploration of course structure and available resources, see the archived course at https://6002x.mitx.mit.edu/.

Parsing tracking logs. Analysis of tracking logs is an established means for understanding student behavior in blended and online courses.[5,14] In the 6.002x tracking logs, each interaction (click) contained relevant information, including username, resource ID, interaction details, and timestamp. Interaction details are context-dependent (such as correctness of a homework problem submission, body text of a discussion post, and page number for book navigation). The edX software is distributed through the cloud; meaning interaction data is logged on multiple servers. In total, approximately 230 million interactions were logged in 38,000 log files over the initial Spring 2012 semester.

We preprocessed the logs into separate time-series for each participant, then compiled participant-level descriptive statistics on resource usage, including number of unique resources accessed, total frequency of accesses per resource type, and total time spent per resource. We also parsed problem submissions, generating a response matrix including correctness and number of attempts. Where possible, we crosschecked our event-log assessment data against a MySQL database serving the 6.002x courseware. All log parsing was performed through standard modules in Python and R.

Estimation of time spent on resources. Time estimation for each participant involved measuring the durations between a student's initial interaction with a resource and the time the student would navigate away. We accumulated durations calculated from each participant's time series for each separate course component, including homework, book, and discussion forums. We found evidence that durations shorter than three seconds represent students navigating to desired resources; hence, we do not count these intervals as activity. In addition, we did not accumulate durations longer than one hour, assuming users have disengaged from their computers. Using alternate values of the high cutoff (20 minutes to one hour) can change overall time by 10%–20% but did not significantly alter relationships regarding time allocation among course components or total time spent by different participants.

An important point is that time accumulated is associated with the resource displayed at the moment; for example, if a student references the book while doing homework, this duration is accumulated with book time. In our case,

only direct interactions with the homework are logged with homework resources. There are clearly alternatives to this approach (such as considering all time between opening and answering a problem as problem-solving time[21]). Our time-accumulation algorithm is partially thwarted by users who open multiple browser windows or tabs; edX developers are considering ways to account for this in the future.

Results

The novelty and publicity surrounding MOOCs in early 2012 attracted a large number of registrants who were more curious than serious. We still take participation in assessment as an indication of serious intent. Of the 154,000 registrants in 6.002x in spring 2012, 46,000 never accessed the course, and the median time spent by all remaining participants was only one hour (see Figure 2a). We had expected a bimodal distribution of total time spent, with a large peak of "browsers" who spent only on the order of one hour and another peak from the certificate earners at somewhere more than 50 hours. There was, in fact, no minimum between

Figure 1. Screenshot of typical student view in 6.002x.

All course components are accessed from the interface shown below. The left sidebar defines the course sequence; weekly units include lecture sequences (videos and questions), homework, lab, and tutorials. The header navigation provides access to supplementary materials, including digital textbook, discussion forums, and wiki. The main frame represents the first lecture sequence; beige boxes below the header indicate lecture videos and questions.

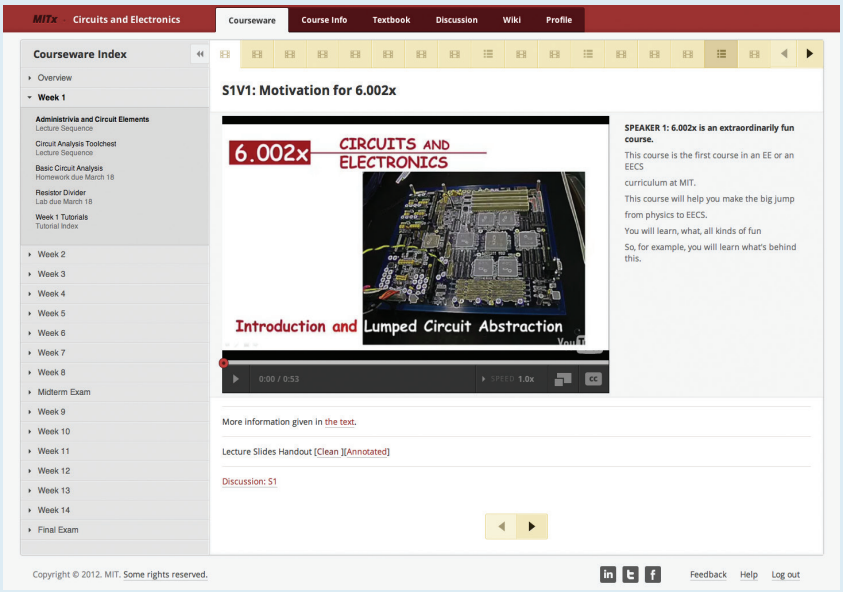


Figure 2. Tranches, total time, and attrition.

(a) Distribution of time spent by participants in 6.002x (time axis is log-transformed); we divided the noncertificate earners into tranches based on percentage of assessment activity they attempted (see also Table 1); (b) percentage of total measured time spent by each tranche; and (c) average time a student invested per week. The shaded regions near Week 8 and Week 14 represent the time span for the midterm and final exams.

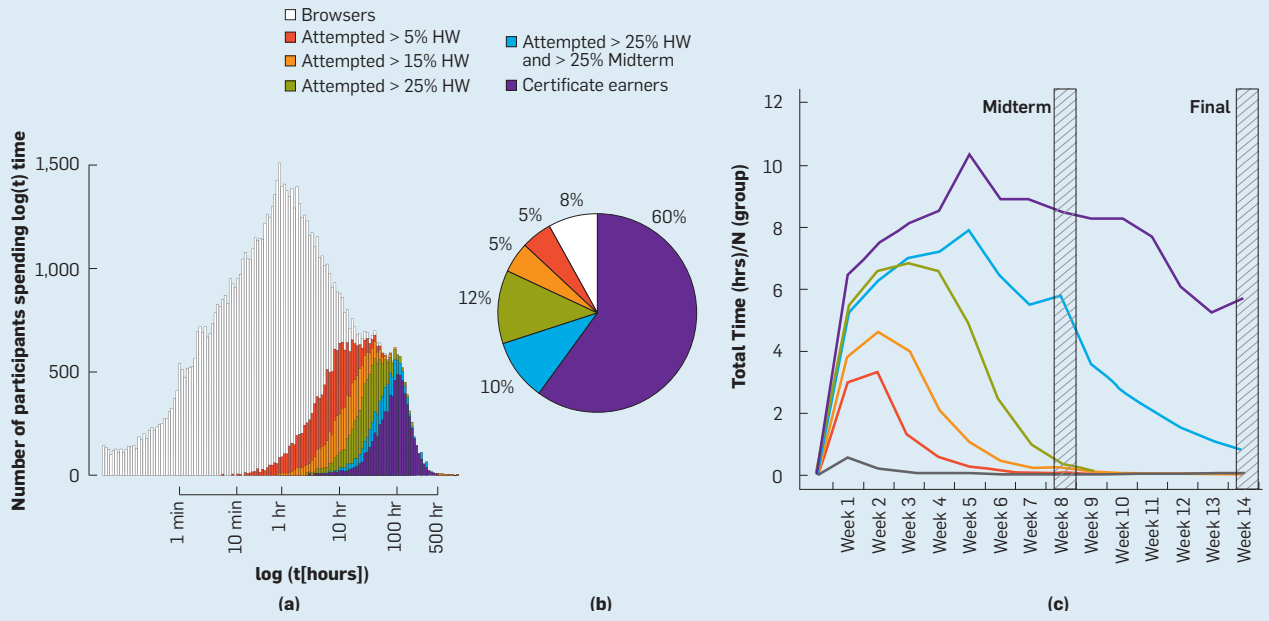
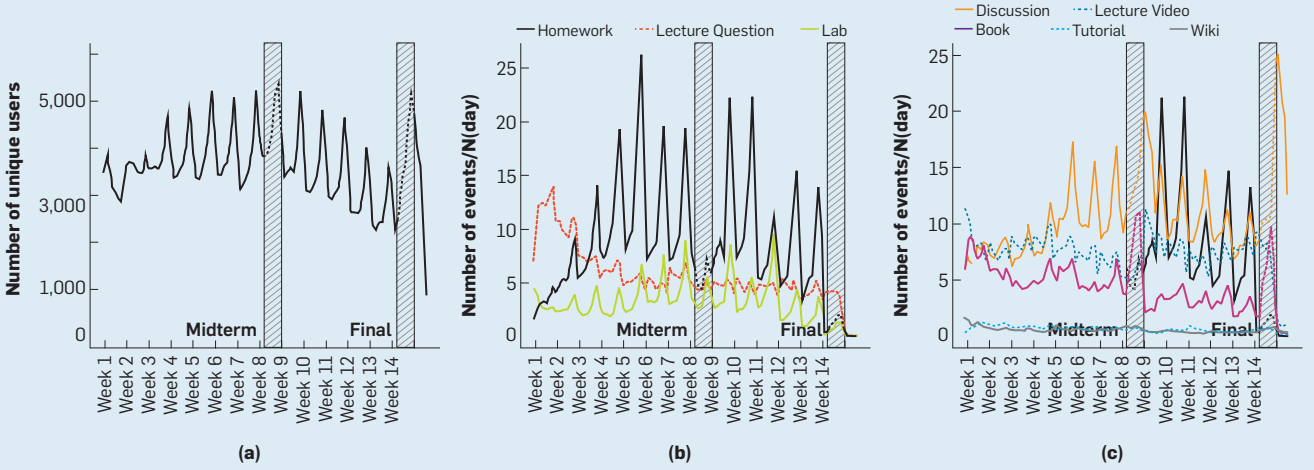


Figure 3. Frequency of accesses.

From left to right, number of unique certificate earners N active per day, their average number of accesses each day for assessment-based and learning-based course components. Plot (a) highlights the periodicity and trends of the certificate earners. Plot (b) is for assessment, including homework, lab, and lecture questions, showing number of accesses per active users that day. Learning-based components in plot (c) include lecture videos, textbook, discussion, tutorial, and wiki, showing discussion forums were used more heavily and with strong periodicity later in the term, similar to graded activities in plot (a), while other components lack periodicity and vary greatly in terms of frequency of accesses. The shaded regions near Week 8 and Week 14 represent the time span for the midterm and final exams.



these extremes, only a noticeable shoulder (see Figure 2a). The intermediate durations are filled with attempters we divided into tranches (in colors) on the basis of how many assessment items they attempted on homework and exams: browsers (gray) attempted < 5% of homework; tranche 1 (red) 5%–15% of homework; tranche 2 (orange) 15%–25% of homework; tranche 3 (green) > 25% of homework; and tranche 4 (cyan) >25% of homework and 25% of midterm exam. Certificate earners (purple) attempted most of the available homework, midterm, and final exams. The median total time spent in the course for each tranche was 0.4 hours, 6.4 hours, 13.1 hours, 30.0 hours, 53.0 hours, and 95.1 hours, respectively. In addition to these tranches, just over 150 certificate earners spent fewer than 10 hours in the course, possibly representing a highly skilled tranche seeking certification. Similarly, just over 250 test takers spent fewer than 10 hours in the course and completed more than 25% of both exams but did not earn a certificate.

The average time spent in hours per week for participants in each tranche is shown in Figure 2c. Tranches attempting fewer assessment items not only taper off earlier, as the majority of participants effectively drop out, but also invested less time in the first few weeks than the certificate earners. The correlation of attrition with less time spent in early weeks begs the question of whether motivating students to invest more time would increase retention rates.

In the rest of this article, we restrict ourselves to certificate earners, as they accounted for the majority of resource consumption; we also wanted to study time and resource use over the whole semester.

Frequency of accesses. Figure 3a shows the number of active users per day for certificate earners, with large peaks on Sunday deadlines for graded homework and labs but not for lecture questions. There is a downward trend in the weeks between the midterm and the final exam (shaded regions). No homework or labs were assigned in the last two weeks before the final exam, though the peaks persist. We plotted activity in events (clicks subject to time cutoffs) per active student per day for assessment-based course components and learning-based components in Figure 3b and Figure 3c. Homework sets and the discussion forums account for the highest rate of activity per student, with discussion activity increasing over the semester. Lecture question events decay early as homework activity increases. Textbook use peaks during exams, and there is a noticeable drop in textbook activity after the midterm, as is typical in traditional courses.[18]

Time on tasks. Time represents the principal cost function for students, so it is important to study how students allocate time among available course components.[15,19] Figure 4 shows the most time is spent on lecture videos; since three to four hours per week is close to the total duration of the scheduled videos, students who rewound and reviewed the videos must compensate for those speeding up playback or omitting videos.

The most significant change over the first seven weeks was the apparent transfer of time from lecture questions to homework, as in Figure 4. Considering a performance-goal orientation (see Figure 5), it should be noted that homework counted toward the course grade, whereas lecture questions did not. But even on mastery-oriented grounds, students might have viewed completion of homework as sufficient evidence of understanding lecture content. The prominence of time spent in discussion

forums is especially noteworthy, as they were neither part of the course sequence nor did they count for credit. Students presumably spent time in discussion forums due to their utility, whether pedagogical or social or both. The small spike in textbook time at the midterm, a larger peak in the number of accesses, as in Figure 3, and the decrease in textbook use after the midterm are typical of textbook use when online resources are blended with traditional on-campus courses.[18] Further studies comparing blended and online textbook use are also relevant.[3,17]

Percentage use of course components. Along with student time allocation, the fractional use of the various course components continues to be an important metric for instructors deciding how to improve their courses and researchers studying the influence of course structure on student activity and learning. For fractional use, we plotted the percentage of certificate earners having accessed at least a certain percentage of resources in a course component (see Figure 5). Homework and labs (each 15% of overall grade) reflect high fractional use. The inflection in these curves near 80% might have been higher but for the course policy of dropping the two lowest-graded assignments. The low proportionate use of textbook and tutorials is similar to the distribution observed for supplementary (not explicitly included in the course sequence) e-texts in large introductory physics courses,[16] though the 6.002x textbook was assigned in the course syllabus. The course authors were disappointed with the limited use of tutorial videos, suspecting that placing tutorials after the homework and laboratory (they were meant to help) in the course sequence was partly responsible. (The wiki and discussion forums had no defined number of resources so are excluded here.)

To better understand the middle curves representing lecture videos and lecture problems, it helps to recall that the negative slope of the curve is the density of students accessing that fraction of that course component (see Figure 5b and Figure 5c). Interestingly,

Figure 4. Time on tasks.

Certificate earners average time spent, in hours per week, on each course component; midterm and final exam weeks are shaded.

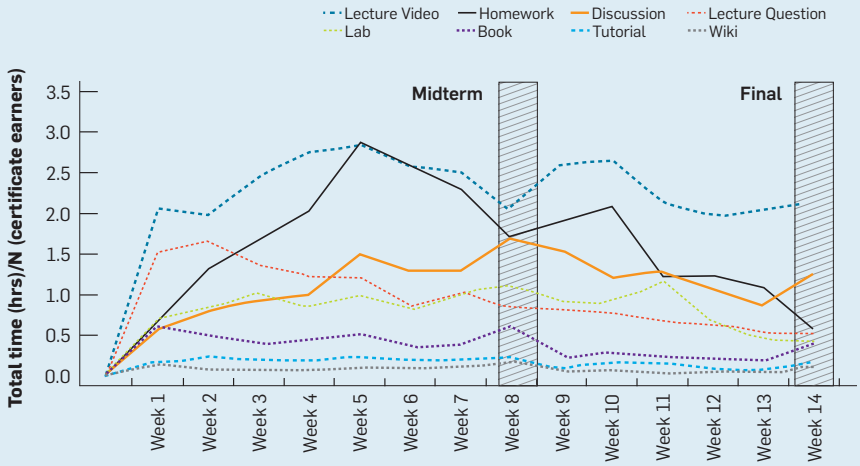
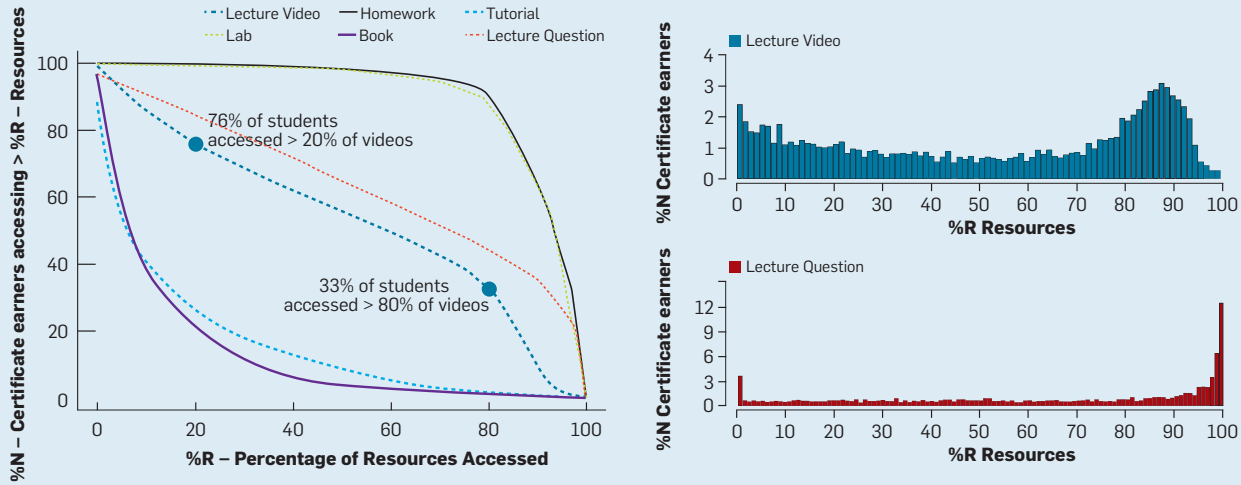


Figure 5. Fractional use of resources.

(a) Percentage of certificate earners who accessed greater than %*R* of that type of course resource. The density of users is the negative slope of the usage curve. Two points indicating bimodality of lecture video use are plotted: 76% of students accessed > 20% of lecture videos, and 33% of students accessed > 80% of lecture videos. (b) Bimodal distribution for videos accessed (as percentage). And (c) distribution of lecture questions accessed.



the distribution for the lecture videos is distinctly bimodal: 76% of students accessed over 20% of the videos (or 24% of students accessed less than 20%), and 33% accessed over 80% of the videos. This bimodality merits further study into learning preferences; for example, do some students learn from other resources exclusively? Or did they master the content prior to the course? The distribution of lecture-problem use is flat between 0% and 80%, then rises sharply, indicating that many students accessed nearly all of them. Along with the fact that the time on lecture questions drops steadily in the first half of the term (see Figure 2), this distribution suggests students not only allocated less time to them, some abandoned the lecture problems entirely.

Resources used when problem solving. Patterns in the sequential use of resources by students may hold clues to cognitive and even affective state.[2] We therefore explored the interplay between use of assessment and learning resources by transforming time-series data into transition matrices between resources. The transition matrix contains all individual resource-resource transitions we aggregated into transitions between major course components. The completeness of the 6.002x learning environment means students did not have to leave it to reference the textbook, review earlier homework, or search the discussion forums. We thus had a unique opportunity to observe transitions to all course components accessed by students while working problems. In previous studies of online problem solving this information was simply missing.[21]

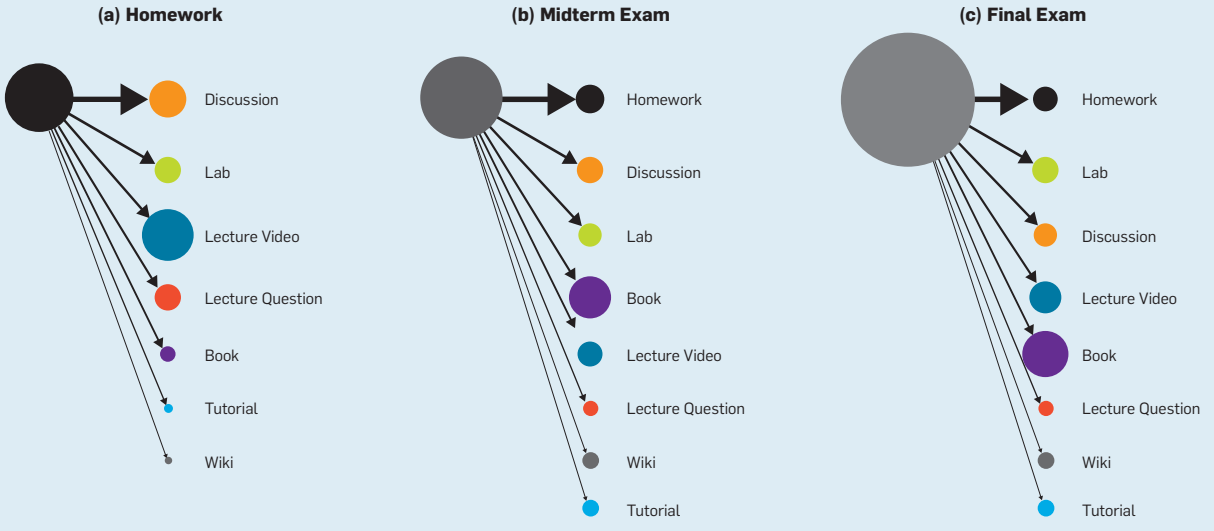
Figure 6 highlights student transitions from problems (while solving them) to other course components, treating homework sets, the midterm, and the final exam as separate assessment types of interest. Figure 6 shows the discussion forum is the most frequent destination during homework problem solving, though lecture videos consume the most time. During exams (midterm and final are similar), previously done homework is the primary destination, while the book consumes the most time. Student behavior on exam problems thus contrasts sharply with behavior on homework problems. Note that because homework was aggregated, we could not isolate "references to previous assignments" for students doing homework.

Conclusion

This article's major contribution to course analysis is showing how MOOC data can be analyzed in qualitatively different ways to address important issues: attrition/retention, distribution of students' time among resources, fractional use of those resources, and use of resources during problem solving. Among the more significant findings is that participants who attempted over 5% of the homework represented only 25% of all participants but accounted for 92% of the total time spent in the course; indeed, 60% of the time was invested by the 6% who ultimately received certificates. Participants who left the course invested less effort than certificate earners, with those investing the least effort during the first two weeks tending to leave sooner. Most certificate earners invested the plurality of their time in lecture videos, though approximately 25% of the earners watched less than 20%. This suggests the need for a follow-up investigation into the correlations between resource use and learning. Finally, we highlight the significant popularity of the discussion forums in spite of being neither required nor included in the navigation sequence. If this social learning component played a significant role in the success of 6.002x, a totally asynchronous alternative might be less appealing, at least for a complex topic like circuits and electronics.

Some of these results echo effects seen in on-campus studies of how course structure affects resource use[18] and performance outcomes[4,11,19] in introductory (college) courses. This

Figure 6. Transitions to other components during problem solving on (a) homework, (b) midterm, and (c) final. Arrows are thicker in proportion to overall number of transitions, sorting components from top to bottom; node size represents total time spent on that component.



and future MOOC studies should further illuminate on-campus education generally. On the other hand, MOOCs could well take advantage of insights from existing research in on-campus education (such as frequent exams drive resource use and maximize learning outcomes[11]).

Finally, we emphasize that MOOCs provide a unique view into the learning of a large, diverse population of students, allowing research based on detailed insight into all aspects of a course. In contrast to most previous studies of on-campus educational environments, we have time-stamped logs of essentially all student behavior and associated learning throughout the entirety of a course, all with solid statistics and the ability to study specific student cohorts (such as based on effort, learning habits, and demographics[9]). Combining time-on-task observations with measures of learning paves the way for measuring learning value—the amount learned per unit time spent on a given course component—possibly extending previous studies of online learning.[7,15] This, in turn, will allow a process of cyclic improvement based on research development, experimentation, and measurement of learning outcomes, supporting improvement of educational content and delivery. Since many MOOCs largely mirror traditional on-campus courses in types of resources, format, and chronology, we anticipate insights into, and improvements of, learning in traditional on-campus courses as well.

Acknowledgments

This work is supported, but is not endorsed, by National Science Foundation grant DUE-1044294; additional support provided by a Google Faculty Award. We thank MITx for data access and J. deBoer and other members of the Teaching and Learning Laboratory and the Research in Learning, Assessing and Tutoring Effectively groups at MIT for their helpful suggestions and comments. C

Textbook use peaks during exams, and there is a noticeable drop in textbook activity after the midterm, as is typical in traditional courses.

References

1. Ames, C. and Archer, J. Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology 80*, 3 (1988), 260.
2. Baker, R.S., D'Mello, S.K., Rodrigo, M.M.T., and Graesser, A.C. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies 68*, 4 (2010).
3. Cummings, K., French, T., and Cooney, P.J. Student textbook use in introductory physics. In *Proceedings of the Physics Education Research Conference*, 2002.
4. Freeman, S., Haak, D., and Wenderoth, M.P. Increased course structure improves performance in introductory biology. *CBE-Life Sciences Education 10*, 2 (2011).
5. Guzdial, M. *Deriving Software Usage Patterns from Log Files.* Technical Report GIT-GVU-93-41, 1993.
6. Hyman, P. In the year of disruptive education. *Commun. ACM, 55*, 12 (Dec. 2012).
7. Jiang, L., Elen, J., and Clarebout, G. The relationships between learner variables, tool-usage behavior, and performance. *Computers in Human Behavior 25*, 2 (2009).
8. Johnstone, S.M. Open educational resources serve the world. *Educause Quarterly 28*, 3 (2005).
9. Kolowich, S. Who takes MOOCs? *Inside Higher Education 5* (2012).
10. Kortemeyer, G. Gender differences in the use of an online homework system in an introductory physics course. *Physical Review Special Topics: Physics Education Research 5*, 1 (2009).
11. Laverty, J.T., Bauer, W., Kortemeyer, G., and Westfall, G. Want to reduce guessing and cheating while making students happier? Give more exams! *Physics Teacher 50*, 9 (2012).
12. Martin, F.G. Will massive open online courses change how we teach? *Commun. ACM 55*, 8 (2012).
13. McAuley, A., Stewart, B., Siemens, G., and Cormier, D. The MOOC model for digital practice. Social Sciences and Humanities Research Council, *Knowledge Synthesis Grant on the Digital Economy*, 2010.
14. Minaei-Bidgoli, B., Kortemeyer, G., and Punch, W.F. Enhancing online learning performance: An application of data mining methods. *Immunohematology 62*, 150 (2004).
15. Morote, E.S. and Pritchard, D.E. What course elements correlate with improvement on tests in introductory Newtonian mechanics? *American Journal of Physics 77* (2009).
16. Pappano, L. The year of the MOOC. *The New York Times* (Nov. 2, 2012).
17. Podolefsky, N. and Finkelstein, N. The perceived value of college physics textbooks: Students and instructors may not see eye to eye. *The Physics Teacher 44* (2006).
18. Seaton, D.T., Bergner, Y., Kortemeyer, G., Rayyan, S., Chuang, I., and Pritchard, D.E. The impact of course structure on etext use in large-lecture introductory-physics courses. In *Proceedings of the Physics Education Research Conference*, 2013.
19. Stewart, J., Stewart, G., and Taylor, J. Using time-on-task measurements to understand student performance in a physics class: A four-year study. *Physical Review Special Topics-Physics Education Research 8*, 1 (2012).
20. Vardi, M.Y. Will MOOCs destroy academia? *Commun. ACM 55*, 11 (Nov. 2012).
21. Warnakulasooriya, R., Palazzo, D.J., and Pritchard, D.E. Time to completion of Web-based physics problems with tutoring. *Journal of the Experimental Analysis of Behavior 88*, 1 (2007).

Daniel T. Seaton (dseaton@mit.edu) is a postdoctoral research fellow in the Office of Digital Learning at the Massachusetts Institute of Technology, Cambridge, MA.

Yoav Bergner (ybergner@ets.org) is a research scientist in the Center for Advanced Psychometrics at ETS, Princeton, NJ.

Isaac Chuang (ichuang@mit.edu) is a joint professor in the Department of Physics and the Department of Electrical Engineering and Computer Science and a member of the Research Laboratory of Electronics at the Massachusetts Institute of Technology, Cambridge, MA.

Piotr Mitros (piotr@mitros.org) is the chief scientist at edX and affiliated with the Center for Artificial Intelligence and Learning at the Massachusetts Institute of Technology, Cambridge, MA.

David E. Pritchard (dpritch@mit.edu) is the Cecil and Ida Green Professor of Physics and a member of the Center for Ultracold Atoms and the Research Laboratory for Electronics at the Massachusetts Institute of Technology, Cambridge, MA.

Copyright held by Author(s)/Owner(s)