Who does what in a massive open online course?

Who Does What in a Massive Open Online Course?

Article *in* Communications of the ACM · September 2013

DOI: 10.1145/2500876

5 authors, including:

Daniel T. Seaton
Harvard University
54 PUBLICATIONS **970** CITATIONS
SEE PROFILE

Yoav Bergner
Educational Testing Service
15 PUBLICATIONS **208** CITATIONS
SEE PROFILE

Piotr Mitros
Massachusetts Institute of Technology
16 PUBLICATIONS **218** CITATIONS
SEE PROFILE

David E. Pritchard
Massachusetts Institute of Technology
414 PUBLICATIONS **17,248** CITATIONS
SEE PROFILE

Title: Who Does What in a Massive Open Online Course?

Preprint – Accepted by Communications of the ACM

Authors:

Daniel T. Seaton[1,2]*, Yoav Bergner[2], Isaac Chuang[1-3], Piotr Mitros[4], and David E. Pritchard[2]

Affiliations:

[1]Office of Digital Learning, Massachusetts Institute of Technology, Cambridge, MA 02139.

[2]Department of Physics and Research Laboratory for Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139.

[3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

[4]edX and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139.

*Correspondence to: dseaton@mit.edu.

Abstract:

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. We present an overview of how the 108,000 participants behaved in 6.002x - *Circuits and Electronics*, the first course in MITx (now edX). We divide participants into tranches based on the extent of their assessment activities, ranging from browsers (who constituted ~76% of the participants but accounted for only 8% of the total time spent in the course) to certificate-earners (7% of the participants who accounted for 60% of the total time). We examine how the certificate earners allocated their time amongst the various course components and study 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 use of various course components, and transitions among them, influence learning in MOOCs.

One Sentence Summary: We analyze learner behavior in the inaugural edX course (6.002x: Circuits and Electronics), including participation level, instructional resource usage, and time allocation.

Introduction:

Though free online courses are not new [8], they have reached an unprecedented scale since late 2011. Three organizations (Udacity, Coursera, and edX) released massive open online courses (MOOCs) [13] that drew over 100,000 registrants per course. In the year since, numbers from these three initiatives have grown to over 100 courses and 3 million total registrants, resulting in 2012 being dubbed "The Year of the MOOC" by the New York Times [16]. Although there has

been much speculation regarding how these most recent MOOC initiatives may reshape higher education [6,12,20], to date, little analysis has been published describing student behavior or learning in these MOOCs.

The main objective of this article is to show that the huge amount of data available in MOOCs offers a unique research opportunity: a means to study detailed student behavior in a complete self-contained learning environment throughout an entire course. We study the approximately 100GB of time-stamped log data describing student interactions with the inaugural MITx (now edX) course, 6.002x *Circuits and Electronics*, in Spring 2012: data at least two orders of magnitude larger than analyzed in previous studies of online learning [21,10]. We develop and exhibit several ways to study student interactions with course resources. We do not analyze demographic factors, but rather differentiate students by the number of assessment items attempted and by the total time spent in the course. All registrants are studied with these metrics before turning our attention to the more detailed time allocation and resource usage of students who earned a certificate of accomplishment. For certificate-earners, we examine the use of different course components (e.g. lecture videos, homework, discussion forum, etc.) in terms of user time allocation and the total fraction accessed. We also study 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.

6.002x, Procedures, and Data Analysis:

With some modification for online delivery, the 14 weeklong units of 6.002x largely mirror a traditional on-campus course in both format and timing. The course sequence (left navigation bar in Fig 1) comprises lecture sequences consisting of lecture videos (annotated powerpoints and actual MIT lectures) with embedded lecture questions, tutorial videos (recitation substitute), homework (3-4 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 (top navigation bar in Fig. 1) include a course textbook (navigable page images), a TA- and student-editable wiki, and moderated student discussions. For further exploration of course structure and available resources, readers may visit the archived course (https://6002x.mitx.mit.edu/).

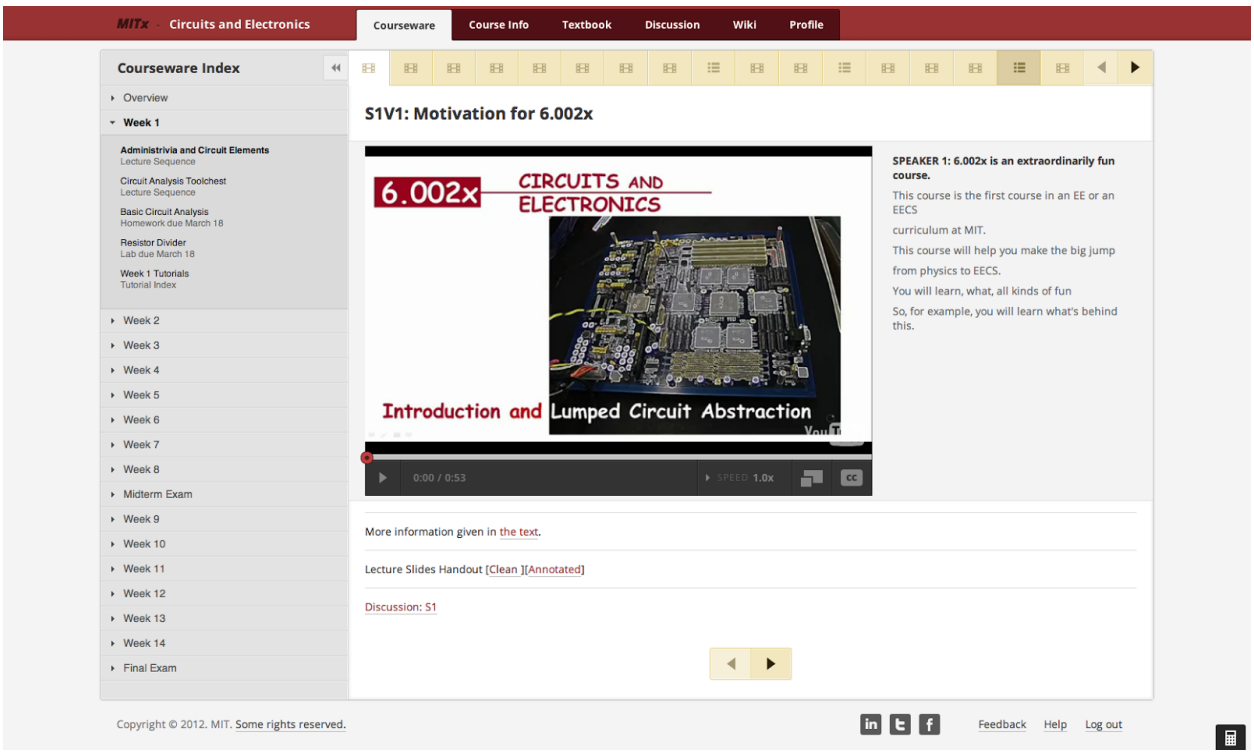

Fig. 1: Screenshot of typical student view in 6.002x. All course components are accessed from this interface. The left sidebar defines the *course sequence*; weekly units contain lecture sequences (videos and questions), homework, lab, and tutorials. The header navigation provides access to supplementary materials: digital textbook, discussion forums, and wiki. The main frame represents the first lecture sequence; beige boxes below the header indicate lecture videos and questions.

Parsing tracking logs

Analysis of tracking logs is a well-established means for analyzing student behavior in blended and online courses [5,14]. In the 6.002x tracking logs, each interaction (*click*) contains the following relevant information: *username, resource id, interaction details,* and a *timestamp.* Interaction details are context-dependent, e.g., correctness of a homework problem submission, body text of a discussion post, page number for book navigation. The edX software is distributed in the cloud; meaning interaction data are logged on multiple servers. In total, roughly 230 million interactions were logged in 38,000 log files.

We preprocessed the logs into separate time-series for each participant, then compiled participant-level descriptive statistics on resource usage: number of unique resources accessed, total frequency of accesses per resource type, and the total time spent per resource. Additionally, we parsed problem submissions, generating a response matrix that includes correctness and the 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 using standard modules in Python and R.

Estimation of time spent on different resources

Time estimation for each participant involves measuring the durations between a student's initial interaction with a resource and the time when they navigate away. We accumulate durations calculated from each participant's time-series for each separate course component (Homework, Book, Discussion Forums, etc.). We have evidence that those durations less than 3 seconds represent students navigating to desired resources, hence, we don't count these intervals as activity. In addition, we don't accumulate durations over 1 hour, assuming that the user has disengaged from their computer. Using alternate values of the high cutoff (20 min to 1 hr) can change overall times by 10-20%, but does 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 currently displayed resource. For example, if a student references the book while working on the homework, this duration is accumulated with book time. Only direct interactions with the homework are logged with homework resources. There are clearly alternatives to this approach, e.g. 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:

From Browsers to Certificate Earners:

The novelty and publicity surrounding MOOCs in early 2012 attracted a large number of registrants who were more curious than serious. We take participation in assessment as an indication of serious intent. Of the 154k registrants in 6.002x, 46k never accessed the course, and the median time spent by all remaining participants was only 1 hour (see Fig. 2A). We had expected a bimodal distribution of total time spent, with a large peak of "browsers" who spent only on the order of an hour and another peak from the certificate earners at somewhere over 50 hours. In fact, there was no minimum between these extremes, only a noticeable shoulder, see Fig. 2A. The intermediate durations are filled with attempters whom we divide into tranches (shown 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, *Tranche* 4 (cyan) >25% of homework and 25% of midterm exam. Certificate earners (purple) attempted a majority of homework, midterm and final exams. The median total time spent in course for each tranche was 0.4 hrs, 6.4 hrs, 13.1 hrs, 30.0 hrs, 53.0 hrs, and 95.1 hrs, respectively. In addition to these tranches, there were just over 150 certificate earners that spent less than 10 hrs in the course, possibly representing a highly skilled tranche seeking certification. Similarly, there were just over 250 *test-takers* who spent less than 10 hrs in the course and completed greater than 25% of both exams, but did not earn a certificate.

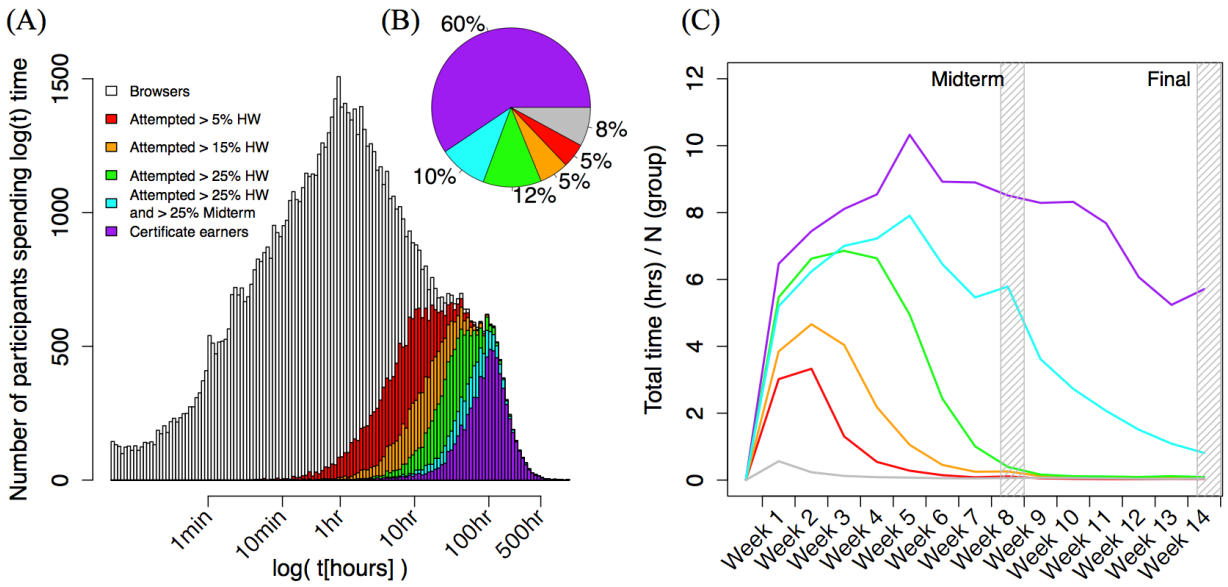

Fig. 2: Tranches, Total time, and Attrition. (A) Distribution of time spent by participants in the course (time axis is log-transformed). We have divided the non-certificate earners into tranches based on the percentage of assessment activity they attempted as shown in Table I. (B) Percentage of total measured time spent by each tranche. (C) Average time a student in various tranches invests per week.

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

In the remainder of this paper, we restrict our attention to the certificate-earners both because this group accounted for the majority of resource consumption, but also because we want to study time and resource use over the whole semester.

Frequency of accesses

Fig. 3A shows the number of active users per day for certificate earners, where large peaks occur 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 plot activity in events (clicks subject to time cutoffs) per active student per day for assessment-based course components and learning-based components in Fig. 3B and 3C. Homework sets and the discussion forums account for the highest activity per student, with discussion activity increasing over the semester - apparently as students use it as a resource when doing homework, see Fig 6. 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].

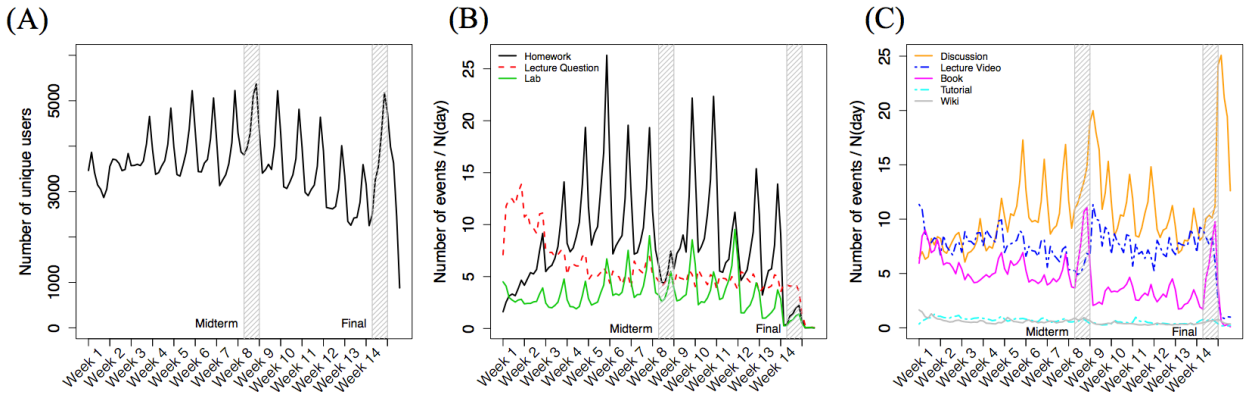

Fig. 3: Frequency of accesses. From left to right, number of unique certificate earners N active per day, their average number of accesses each day for assessment-based (middle) and learning-based course components (right panel) each day. Plot (A) highlights the periodicity and trends of the certificate earners. Plot (B) (for assessment: Homework, Lab, Lecture questions) shows the number of accesses per active users that day. Learning-based components in plot (C) (Lecture videos, Textbook, Discussion, Tutorial, Wiki) show that 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 the frequency of accesses.

Time on Tasks

Time represents the principal cost function for students and it is therefore important to study how students allocate time among available course components [15, 19]. Figure 4 shows that the most time is spent on lecture videos; since 3-4 hours per week is close to the total duration of the scheduled videos, students who rewind and review the videos must compensate for those who speed-up playback or omit videos.

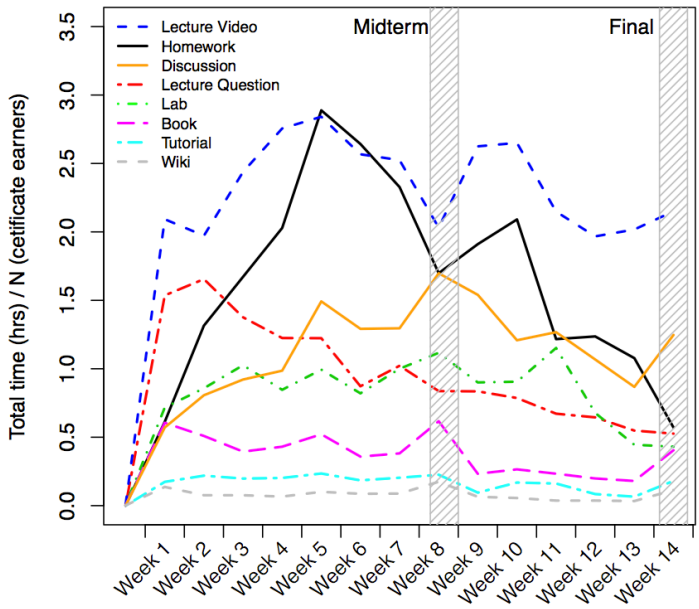

Fig. 4: Time on tasks. Certificate earners average time spent in hours per week on each course component. Midterm and final exam weeks are shaded.

The biggest change over the first seven weeks is the apparent transfer of time from the lecture questions to the homework (see Fig. 4). Considering a performance goal orientation (5), it should be noted that homework counted towards the course grade, whereas lecture questions did not. But even on mastery-oriented grounds, it is possible that students saw completion of the homework as sufficient evidence of understanding the lecture content. The prominence of time spent in the discussion forums is noteworthy as these were neither part of the course sequence, nor did they count for credit. Presumably students spent time in the discussion forums because of 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 (see Fig. 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 usage of the various course components is an important metric for instructors deciding how to improve their courses and for researchers studying the influence of course structure on student activity and learning. For fractional usage (Fig. 5), we plot the percentage of certificate earners having accessed at least a certain percentage of resources in a course component. Homework and labs (each 15% of overall grade) display high fractional usage. The inflection in these curves around 80% might even be higher but for the course policy of dropping the two lowest assignments. The low proportionate usage of textbook and tutorials is similar to the distribution observed for supplementary (that is, 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 low usage of tutorial videos and suspected that placing tutorials after the homework and laboratory (which they were meant to help) in the course sequence was partly responsible. (The wiki and discussion forums had no defined number of resources and so are excluded here.)

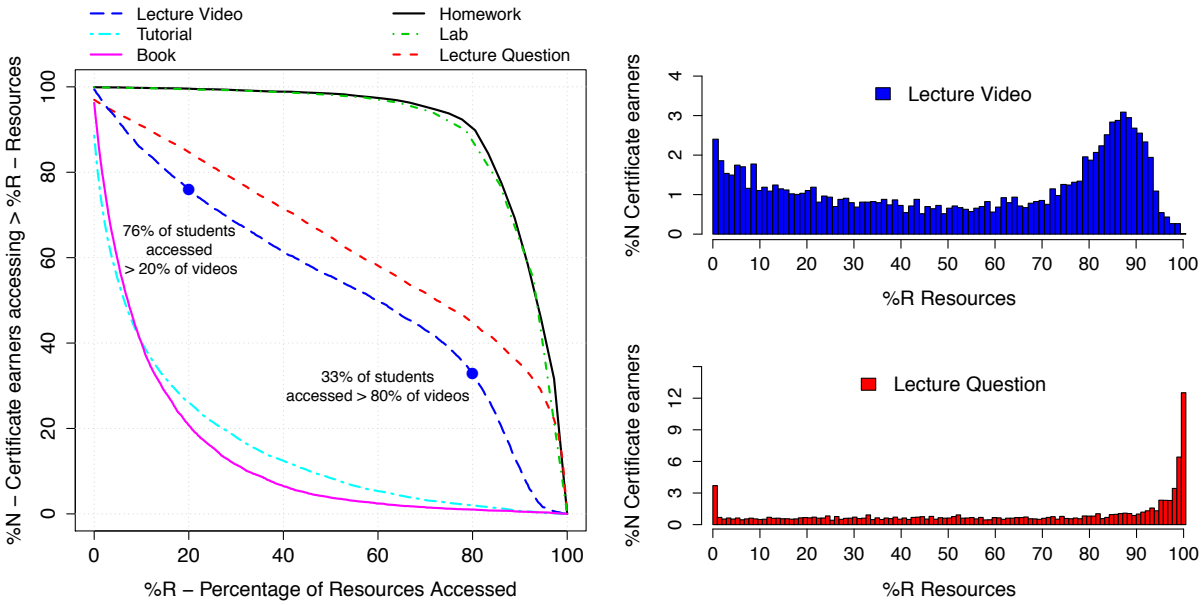

Fig. 5: Fractional Usage of Resources. (A) The percentage of certificate earners that accessed greater than %R of that type of course resource. The density of users is the negative slope of the usage curve. Two points indicating bimodality of Lecture Video use are plotted: 76% of students accessed > 20% of Lecture Videos and 33% of students accessed > 80% of Lecture Videos. (B) The bimodal distribution for videos accessed (as percentage). (C) Distribution for the lecture questions.

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 Fig. 5b and 5c). Interestingly, the distribution for the lecture videos is distinctly bimodal: 76% of students accessed greater than 20% of the videos (or 24% of students accessed less than 20%) and 33% accessed greater than 80% of the videos. This bimodality merits further study into learning preferences: are some students learning from other resources exclusively, or are these participants who have already mastered the content prior to this course? The distribution of lecture problem use is flat between 0% and 80%, and then rises sharply, indicating that many students access nearly all of them. Along with the fact that the time on lecture questions drops steadily in the first half of the term (See Fig. 2), this distribution suggests that students not only allocated less time to them but some abandoned the lecture problems entirely.

Which resources are used while problem solving?

Patterns in the sequential use of resources by the student may hold clues to cognitive and even affective state [2]. Therefore, we have 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, which we aggregate into transitions between major course components. The completeness of the 6.002x learning environment means that students do not have to leave it to reference the textbook, review earlier homework, or search the discussion forums. Hence we have 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]**.

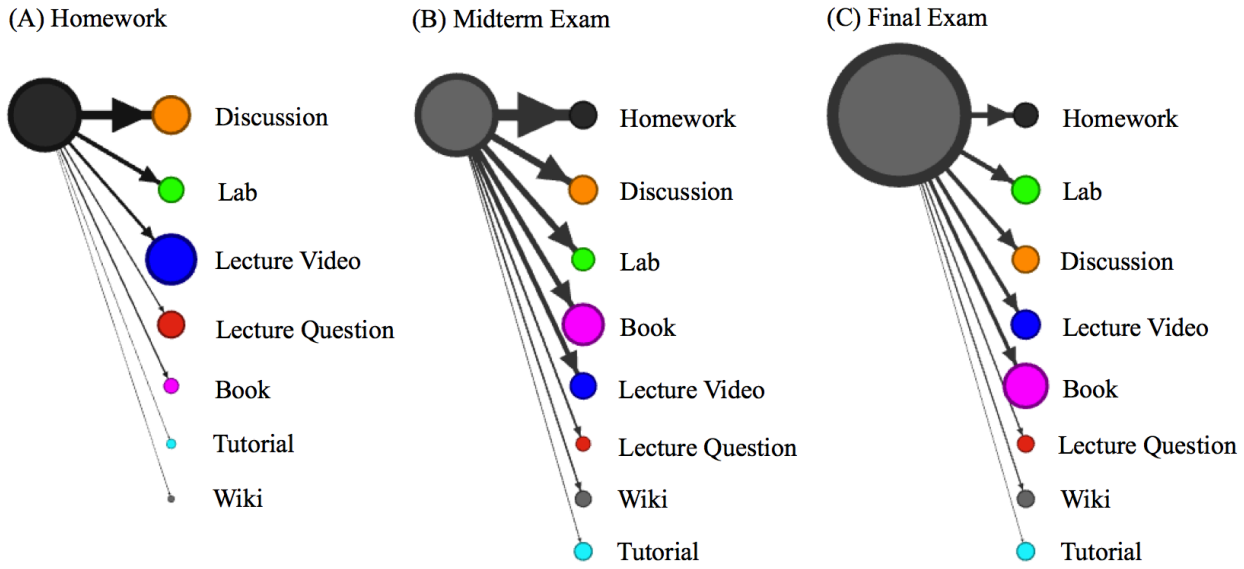

Fig. 6: Transitions to other components during problem solving on the (A) homework, (B) midterm, and (C) final. Arrows are thicker in proportion to the overall number of transitions, which sorts components from top to bottom; node size represents the total time spent on that component.

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 that the discussion forum is the most frequent destination during homework problem solving, while lecture videos consume the most time. During the exams (both midterm and final are similar), previously done homework is the primary destination, while the book consumes the most time. Thus student behavior on exam problems contrasts sharply with behavior on homework problems. Note that because homework were aggregated, we could not isolate "references to previous assignments" for students doing homework.

Conclusions:

The major achievement of this paper is showing how MOOC data can be analyzed in qualitatively different ways to address various issues of importance: attrition/retention, distribution of students' time amongst resources, fractional use of those resources, and use of resources during problem solving. Among the more significant findings are that participants who

attempted over 5% of the homework represented only one-fourth 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 received certificates. Participants who left the course invested less effort than certificate earners, with those who invested the least effort during the first two weeks tending to leave sooner. Most certificate earners invested the plurality of their time in lecture videos; however, approximately one quarter of the earners watched less than 20%. This suggests the need for a follow-up investigation of 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 may be less appealing, at least for a complex topic like circuits and electronics.

Some of the results above 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 and future MOOC studies should further illuminate on-campus education generally. On the other hand, MOOCs could well take advantage of some of insights from existing research on on-campus education (e.g. that frequent exams drive resource usage and maximize learning outcomes [11]).

Finally, we emphasize that MOOCs provide a unique window into understanding the learning of a large, diverse population of students, allowing research based on detailed insight into all aspects of the course. In contrast to most previous studies of on-campus educational environments, we have time-stamped logs of essentially all student behavior and the associated learning throughout the entirety of a course – all with solid statistics and the ability to study specific student cohorts (e.g. based on effort, learning habits, demographics [9], etc.). 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, allowing 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.

References and Notes:

1. Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of educational psychology*, *80*(3), 260.
2. Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). 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), 223-241.
3. Cummings, K., French, T., & Cooney, P. J. (2002). Student textbook use in introductory physics. In *Physics Education Research Conference*.
4. Freeman, S., Haak, D., & Wenderoth, M. P. (2011). Increased course structure improves performance in Introductory Biology. *CBE-Life Sciences Education*, *10*(2), 175-186.

5. Guzdial, M. (1993). Deriving software usage patterns from log files. *Tech Report GIT-GVU-93-41.*
6. Hyman, P. (2012). In the year of disruptive education. *Communications of the ACM*, *55*(12), 20.
7. Jiang, L., Elen, J., & Clarebout, G. (2009). The relationships between learner variables, tool-usage behaviour and performance. *Computers in Human Behavior*, *25*(2), 501-509.
8. Johnstone, S. M. (2005). Open educational resources serve the world. *Educause Quarterly*, *28*(3), 15.
9. Kolowich, S. (2012). Who Takes MOOCs?. *Inside Higher Ed*.
10. Kortemeyer, G. (2009). Gender differences in the use of an online homework system in an introductory physics course. Physical Review Special Topics-Physics Education Research, 5(1), 010107.
11. Laverty, J. T., Bauer, W., Kortemeyer, G., & Westfall, G. (2012). Want to Reduce Guessing and Cheating While Making Students Happier? Give More Exams!. *Physics Teacher*, *50*(9), 540.
12. Martin, F. G. (2012). Will massive open online courses change how we teach? *Communications of the ACM*, *55*(8), 26.
13. McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). The MOOC model for digital practice. *SSHRC Knowledge Synthesis Grant on the Digital Economy*.
14. Minaei-Bidgoli, B., Kortemeyer, G., & Punch, W. F. (2004). Enhancing Online Learning Performance: An Application of Data Mining Methods1. *Immunohematology*, *62*(150), 20-0.
15. Morote, E. S., & Pritchard, D. E. (2009). What course elements correlate with improvement on tests in introductory Newtonian mechanics?. *American Journal of Physics*, *77*, 746.
16. Pappano, L. (2012). The Year of the MOOC. *The New York Times*.
17. Podolefsky, N., & Finkelstein, N. (2006). The perceived value of college physics textbooks: Students and instructors may not see eye to eye. *The Physics Teacher*, *44*, 338.
18. Seaton, D. T., Bergner, Y., Kortemeyer, G., Rayyan, S., Pritchard, D.E. (*In preparation for submission*). The impact of course structure on etext use in blended introductory physics courses?.
19. Stewart, J., Stewart, G., & Taylor, J. (2012). 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), 010114.
20. Vardi, M. Y. (2012). Will MOOCs destroy academia?. *Communications of the ACM*, *55*(11).
21. Warnakulasooriya, R., Palazzo, D. J., & Pritchard, D. E. (2007). Time to completion of web-based physics problems with tutoring. *Journal of the experimental analysis of behavior*, *88*(1), 103.

Acknowledgements:

This work was supported, but is not endorsed, by NSF grant DUE-1044294. We thank MITx for data access, and J. deBoer and other members of the Teaching and Learning Laboratory and RELATE groups at MIT for helpful suggestions and comments.