# Refocusing the Lens on Engagement in MOOCs

**R. Wes Crues, Nigel Bosch, Michelle Perry, Lawrence Angrave, Najmuddin Shaik, and Suma Bhat**
University of Illinois at Urbana-Champaign
crues2, pnb, mperry, angrave, shaik, spbhat2@ illinois.edu

## ABSTRACT

Massive open online courses (MOOCs) continue to see increasing enrollment and adoption by universities, although they are still not fully understood and could perhaps be significantly improved. For example, little is known about the relationships between the ways in which students choose to use MOOCs (e.g., sampling lecture videos, discussing topics with fellow students) and their overall level of engagement with the course, although these relationships are likely key to effective course implementation. In this paper we propose a multilevel definition of student engagement with MOOCs and explore the connections between engagement and students' behaviors across five unique courses. We modeled engagement using ordinal penalized logistic regression with the least absolute shrinkage and selection operator (LASSO), and found several predictors of engagement that were consistent across courses. In particular, we found that discussion activities (e.g., viewing forum posts) were positively related to engagement, whereas other types of student behaviors (e.g., attempting quizzes) were consistently related to less engagement with the course. Finally, we discuss implications of unexpected findings that replicated across courses, future work to explore these implications, and relevance of our findings for MOOC course design.

## ACM Classification Keywords

K.3.1. Computers and Education: Computer Uses in Education; Distance learning.

## Author Keywords

MOOCs, Engagement Patterns, Course Persistence

## INTRODUCTION

Massive open online courses (MOOCs) are known for their informal nature, large enrollments, and wide-reaching audiences. Resulting from this unique combination of attributes, and by virtue of the nature of the learning platform, evaluating whether a student was successful in obtaining knowledge or skills from the MOOC, or was able to attain the goal or intended benefit from participation, is of fundamental importance to instructional designers. It is also true that such an evaluation is vastly different from determining success in formal educational settings and often difficult to ascertain. To address this need, numerous studies have investigated how students interact with the MOOC platform and with other MOOC students, and what learning behaviors can be inferred from their clickstream and other participation data.



*L@S 2018,* June 26–28, 2018, London, United Kingdom



DOI: https://doi.org/10.1145/3231644.3231658

Investigations on dropout (students who initially interact with a MOOC, but then cease interacting before the end of the course) are important to consider. However, not all students who enroll do so with the intention of staying in the course or accessing more than small portions of the course [28]. One such phenomenon has been identified by Clow [7] as the "funnel of participation:" many students hear about the MOOC, then a great deal of them register for the course, but very few have much interaction with the course (either with the content or with other students) after enrolling, and even less so as the course progresses. Realizing that MOOCs are not necessarily formal, some students might access course content to audit or sample course materials [23]; inherently, some students will not fully interact with the course. Because students come to the course with different needs and intentions [36], it may be more advantageous to consider students' longevity in the course, and to investigate their patterns of engagement, than to examine when they drop out–because this positions us to understand better what the course might be offering to them.

As an alternative to predicting dropout in MOOCs, we investigate the problem of modeling *engagement*. Specifically, we consider engagement as a 3-level measure of course activity over time. By predicting what keeps a student engaged in a MOOC, we can uncover which behaviors students exhibit to remain active. The definition of engagement we use is based on two cutoffs of how active a student was over the entire course: we consider whether a student was active for at least 3 weeks, or for 6 weeks out of an 8-week course (representing 37.5% or 75% of the course). In the definition of engagement we explore, we do *not* assume that student activity is linear because it has been shown that students often do not complete a MOOC in the chronological order specified by the instructor or curriculum [15, 11]. Instead, we consider activity throughout the entire duration of the course, allowing for intermittent periods of inactivity.

Toward this end, our goal was to pinpoint which behaviors kept students interacting with the course. Clickstream data, derived from Coursera[1], was used to capture engagement; a student was considered engaged if at least one event was logged during a week in which the course was live. Using background characteristics that students submitted in a survey, along with features derived from their clickstream data as independent variables in an ordinal logistic regression model, we attempted to predict engagement. Because of the large number of features derived from the clickstream, and due to high rates of inactivity and the presumed presence of many different learning goals [11, 28], we employed regularization techniques to estimate these ordinal logistic regression models.

Furthermore, to ensure that our definition of engagement is more widely applicable than to just one course, we analyzed data from five different MOOCs, across different disciplines, which were offered over several iterations, because it has been shown that findings from one MOOC do not necessarily generalize [3], and there are differences in disciplines that might impact engagement [33]. Also, we note that our definition of engagement is flexible and the analytic technique we used is scalable and provides numerically interpretable results. In the remainder of this paper, we present prior work on modeling engagement, our selected modeling technique, specific details of our data and analysis procedures, the findings from modeling our definition of engagement, and a discussion about the implications of our findings.

## RELATED WORK

As mentioned previously, a great deal of work has focused on analyzing and explaining how MOOC students interact in the online system with course materials, other students, and instructors. Much of this work on interaction has typically targeted a single aspect of engagement, and often within a single course. In Table 1 we highlight studies that have used predictive methods to understand which behaviors are indicative of students dropping out of or remaining engaged in MOOCs. For example, some studies have investigated participation in the forums or merely noting activity with the course, but few have considered a host of activities that garner student attention and how this relates to how students remain engaged in the course. Just as wide as the activities used to understand engagement, the definitions of student success, the types of data used to predict success, and the analytic techniques to understand these data are quite variable. In the next subsections, we discuss definitions of student behavioral patterns, the different definitions of success in MOOCs, and, combining these, the types of behaviors that have been indicative of success in MOOCs.

### Behavioral Patterns in MOOCs

With respect to explaining behavior, both Anderson [2] and Kizilcec [23] used students' behavior patterns to identify different types of students. Kizilcec et al. [23] found four different types of students in MOOCs–completers, or those who behave similarly to students in a formal educational setting; auditors, who watch lecture videos but do not complete assignments; disengagers, or those who start the course strong by completing much at the beginning, but little towards the end; and samplers, or those who view at least one, but might view several, video lectures throughout the course. On the other hand, Anderson and colleagues [2] have identified different types of engagement styles of students–viewers (watch lectures but do few assignments), solvers (do many assignments, watch very few lectures), all-rounders (watch lectures and do assignments), collectors (download lectures and do few assignments), and bystanders (have very little activity in the course). Both of these sets of definitions revealed differences between students enrolled in MOOCs.

Both Kizilcec et al. [23] and Anderson et al. [2] shed light on the diversity of students who enroll in MOOCs. Understanding that some students who enroll in MOOCs might participate in the course as a traditional academic student, while others might just be browsing for courses, is helpful in contextualizing student behaviors and understanding which activities might influence students to engage with or to complete the course.

Although dropout, also called stop-out or non-completion, has been relatively consistently defined in the current literature on MOOCs, the definition of succeeding in the course, however, has been more variably defined and we discuss these next.

### Success Indicators

The informal nature of MOOCs, along with the varying goals and motivations students have for enrolling in a MOOC (c.f. [10, 4, 20, 36, 28]), make defining a student's success quite difficult. However, "success" in a MOOC is often measured by *completion*. Completion has often been defined dichotomously, for example whether or not a student has received formal recognition of completing course requirements or has indicated completion through behavioral characteristics. Alternatively, some studies have defined success by tracking how long a student remains active with the course (e.g. [14, 28, 32, 35]).

Several studies have considered indicators of completion in MOOCs that parallel traditional academic markers of success: grades and certificates. When considering students' grades, an average score of 70% or higher on assignments has been used as a proxy for passing the class [8, 9], while others have used a student's total score [18]. These studies analyzed data from clickstream and language produced by students in forums to understand predictors that were associated with passing the classes. Overall, they found active and collaborative students in the forums were more likely to pass the course [8, 9].

Other studies have defined success as a student earning formal recognition of completing the course by earning a certificate. These studies have used students' demographic characteristics [29, 26, 5], how the students view the usefulness of the course [29], quiz attempts or scores [5, 21], activities in forums [26, 21], and engaging with course content [26, 5, 25] to predict whether a student will earn a certificate.

One of these studies [29] found that some words in a students' free-form response to a question about the utility of the course were predictive of students earning a certificate. Others have used features derived from clickstream data and specific activities because background information about students is not as helpful to predict earning a certificate as clickstream activity is

[1] https://www.coursera.org

| Reference | Courses | Prediction Channels | Outcome Measure |
|---|---|---|---|
| de Barba et al. [4] | 1 | Logged activity, survey | Binary completion, final grade |
| Brooks et al. [5] | 1 | Logged activity, demographics | Binary completion |
| Crossley et al. [8] | 1 | Discussion forum text | Binary completion |
| Crues et al. [10] | 1 | Reasons for enrolling, forum activity | Three levels of engagement |
| Greene at al. [14] | 1 | Demographics, survey | Binary completion per week |
| Halawa et al. [16] | 16 | Logged activity | Binary completion |
| He et al. [18] | 1 | Logged activity | Binary pass/fail |
| Jiang et al. [21] | 1 | Logged activity, forum activity, university enrollment | Binary completion, high/low level |
| Kloft et al. [24] | 1 | Logged activity, forum views, country | Binary completion |
| Mullaney and Reich [25] | 1 | Logged activity, forum activity | Binary completion |
| Qiu et al. [26] | 11 | Logged activity, forum activity, demographics | Binary completion |
| Ramesh et al. [27] | 3 | Logged activity, forum activity, forum text | Binary completion |
| Reich [28] | 9 | Logged activity, demographics, survey | Binary completion overall/per week |
| Robinson et al. [29] | 1 | Demographics, survey | Binary completion |
| Sunar et al. [30] | 1 | Forum activity | Binary completion per week |
| Wen et al. [32] | 3 | Forum text | Binary forum participation per week |
| Whitehill et al. [33] | 40 | Logged activity, forum activity, survey, demographics | Binary completion per week |
| Yang et al. [35] | 1 | Forum activity | Binary completion per week |

Table 1: Summary of exemplary studies that included course engagement prediction.

[5]. Additionally, quiz scores [21], along with forum activity and time spent engaging with the course by doing assignments and watching videos [26], were statistically related to earning a certificate in various MOOCs.

Although measures such as grades on assignments and/or earning a certificate are convenient proxies for completion, and certainly indicate success for some students, these rely on the assumption that students intend to be either "completers" [23] or "all-rounders" [2]. In fact, it was found that less than 70% of students enrolled in a MOOC intended to earn a certificate, and, of those, less than 30% of the students who intended to earn a certificate actually did so [29]. To address this potential shortcoming, that completion may not be the best metric of success, we have expanded how we consider student behaviors, beyond dropping out.

### Behaviors Indicative of Success in MOOCs: Predicting Dropout Using Derived Features

Recognizing that a minority of all MOOC students earn certificates, some students are only interested in viewing lectures and not in completing assignments, and students take MOOCs for many different reasons, other measures to identify completion of a course have been developed. These generally take into account the last action logged by the student or whether the student had an extended absence (i.e., lack of activity) in the MOOC. Various features derived from clickstream data, interactions with other students, and survey data have been used to predict if or when a student might drop out.

Clickstream data have been used to explain behavior and predict when a student might stop engaging with the course via the online system. One study defined dropout as watching fewer than half of the videos in the MOOC or being cumulatively absent for four weeks [16]. The study examined students' viewing habits, whether they skipped an assignment, scored poorly on quizzes, and whether the students were lagging behind (viewing lectures two weeks after they were released), to predict whether they would drop out. A different definition of dropout was proposed by Kloft and colleagues [24], where they defined dropout from the clickstream as no activity during a week in the course, followed by no activity during any later week. Here, views of certain pages and viewing videos were used to predict dropout.

Others have sought to predict when a student might stop being an active participant in the course by using forum activity and the text of students' forum posts. Given that active forum participation generally relates to longer persistence or achieving success in MOOCs [10, 26], and because Wen et al. [32] found that students who were engaged and motivated by the content of forum posts were active in the forums longer, it can reasonably be hypothesized that those who show engagement and motivation in the forums are likely to be persistent in the course. More generally, viewing forums has been identified as a relevant predictor of staying engaged throughout all periods of the course, while active forum participation is a strong predictor of remaining in the course during later stages of a MOOC [27]. Beyond measuring engagement and motivation, or activities with the forums, others have found that social positioning within a MOOC shares a relationship with completion of the course [30, 35]. Alongside features derived from clickstream data, others have used survey responses to predict student behavior in MOOCs.

Students' responses to surveys in MOOCs have been used to understand why students enroll in the course, its perceived usefulness, and to assess student backgrounds [10, 14, 22]. In several studies, the responses to surveys have been used with clickstream data to glean insight into who persists in a MOOC. One study asked students whether they stopped participating in a MOOC before the course was completed and found that women and students from Africa, Latin America, and Asia had lower rates of persistence than men and students from other regions, and the students who left did not have enough time to devote to the course [22]. With respect to time, students who intended to devote time to a MOOC were found to have a lower likelihood of dropping out than those who did not [14]. Finally, another study considered students' reasons for enrolling in a MOOC to determine whether these were related to persisting in the MOOC; it was found that the reasons for enrolling did not share a statistically significant relationship with persisting in a MOOC [10].

Regardless of the method and data used to understand why a student might drop out during the course, exploring which behaviors *keep students engaged* with a course has not been the focus of many of the previously described studies. Generally, these studies have posited ways to predict dropout or reveal the reasons that students might not remain engaged in a MOOC. In the next section, we describe a method to examine which behavioral and background characteristics predict our definition of engagement.

## METHODS

### Description of Courses and Data

We used data from five MOOCs that were hosted on Coursera and were created at a large research university in the United States. The five courses were: Creative, Serious and Playful Science of Android Apps (Android); Introductory Organic Chemistry (Ochem); Subsistence Marketplaces (Subsistence); Introduction to Sustainability (Sustainability); and e-Learning Ecologies (Elearning). Each course lasted eight weeks[2]. Following from this, we define engagement for student $i$ as

$$(engagement)_i = \begin{cases} 0 & \text{\# of weeks active } \leq 2 \\ 1 & 3 \leq \text{\# of weeks active } \leq 5 \\ 2 & 6 \leq \text{\# of weeks active } \leq 8 \end{cases} .$$

[2]We note that one offering of the Android MOOC had a 3-week holiday in the middle of the course. For our purposes, although data were obtained during the holiday, we only considered weeks where new content was presented (i.e., we excluded the holiday break when considering weekly activities, and only analyzed the eight weeks in which content was presented).

This definition can be applied to other MOOCs of different lengths by using fractional boundaries (of 0.375 and 0.75) for the course's duration, optionally rounded to the nearest week or appropriate unit of time. We defined this outcome at multiple levels, rather than as binary, and, focusing on what behaviors predict longevity in the course, refer to this outcome as engagement rather than dropout.

For student $i$ to be considered active during week $w$, this student needed only to have any behavior that was logged by the Coursera platform during that week, regardless of the activity that was logged. This could be just one click during the week or participating in multiple activities, and could come in consecutive weeks or sporadically throughout the eight weeks of the course. Given the informal nature of these courses, we took the act of exhibiting any activity during a week of the course to indicate engagement. Table 2 contains longitudinal data with the mean and median number of events logged per student during each week of the course, as well as the distribution for our definition of engagement.

We posit that this definition of engagement can capture relatively subtle information about students enrolled in these MOOCs, particularly students who are not in the course to earn high scores or a certificate. Likewise, some students might not have been interested in some topics of the course, and thus we did not assume that each student had a continuous, linear progression during the course. We decided to use the cut-points of 3 and 6 weeks to capture the difference between students who were interested in only a few lectures, i.e., $engagement = 0$ from those students who were fairly interested in the course and made contact with the course during 3, 4, or 5 weeks, whom we have categorized as $engagement = 1$, and from those who appeared more like a student in a traditional classroom, where $engagement = 2$.

Structurally, the courses had many similarities with respect to their design. Each of the courses had video lectures that presented new material each week, and students either downloaded or streamed these lecture videos, which allowed us to capture students whom Anderson and colleagues described–specifically, the "collectors" [2], who download the lectures versus streaming them. With the exception of the Elearning course, all classes had quizzes that assessed retention of course material, and these quizzes were given weekly. Each course had a forum where students could interact with each other. Additionally, students were asked to complete a survey that contained a few questions, which asked about their demographics. In each course, students could earn a Statement of Accomplishment (SOA), which is a formal recognition of completing course requirements.

Table 3 presents all of the features we used to model engagement. The course-level and survey features were invariant throughout the course. The courses we considered were session-based, meaning they were offered over a specified period of eight weeks, and the section variable denotes in which session of the course the student enrolled. The earned-an-SOA feature denotes whether a student earned an SOA from their activities. In the demographic survey, students were asked to (1) identify their age within seven different age bands, and

| Course | Events by Week | | | | | | | | Engagement Level | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | One | Two | Three | Four | Five | Six | Seven | Eight | 0 | 1 | 2 |
| Android | 10/78.83 | 0/40.05 | 0/27.61 | 0/35.17 | 0/31.92 | 0/23.37 | 0/20.72 | 0/14.77 | 9649 | 3442 | 2774 |
| Ochem | 2/67.07 | 0/52.26 | 0/46.51 | 0/36.58 | 0/31.59 | 0/26.46 | 0/26.01 | 0/17.56 | 1207 | 369 | 569 |
| Subsistence | 3/46.15 | 0/31.84 | 0/29.27 | 0/28.24 | 0/18.15 | 0/15.53 | 0/12.01 | 0/8.67 | 1461 | 437 | 456 |
| Sustainability | 16/74.18 | 8/61.71 | 1/47.25 | 0/42.03 | 0/47.22 | 0/41.6 | 0/45.5 | 0/43.74 | 4802 | 1754 | 3503 |
| Elearning | 3/42.88 | 0/26.67 | 0/17.37 | 0/12.95 | 0/12.64 | 0/11.4 | 0/9.25 | 0/7.64 | 1122 | 395 | 311 |

Table 2: Summary statistics by week for all students. The first number in each week is the median number of events and the second number is the mean number of events. The right side of the table includes counts for the number of students within each level of engagement.

(2) identify their gender. We excluded students who did not provide both gender and age information. Furthermore, we were also given students' time-zone information and nearest major city from their web browser settings; we reduced this information to the student's continent for analysis.

The lecture and quiz activity features captured quantitative information about how students interacted with the lectures and quizzes over the entire course, not during specific weeks. To this end, we used information captured about the length and percentage of lectures students watched over the entire course, as well as whether they had any interaction with the lectures. For the courses that had quizzes, we considered the number and percentage of quizzes attempted, as well as the maximum score a student achieved on a quiz and the student's score on the first quiz[3].

Finally, we included weekly and summary measures of activities in these MOOCs. For each of the eight weeks, we included each student's count of the number of forum views, posts, and comments. We also included a count of the number of days in each of the eight weeks a student was active within the course. This also relates to the summary measures of interaction: the total number of days a student was active and the percent of days that student was active.

Using all of these potential predictors of engagement, the purpose of our analyses was to uncover what behaviors were related to our definition of engagement. This encouraged us to explore methods that take into account the ordinality of engagement.

| | |
|---|---|
| **Course Level Features** | |
| Section | Earned an SOA |
| **Survey Features** | |
| Age Group | Gender |
| Timezone | |
| **Lecture & Quiz Activity** | |
| AnyLectureActivity | TotalDownloadOnly |
| TotalStream | TotalLecturesWatched |
| LengthLecturesWatched | PercentLecturesWatched |
| PercentLengthWatched | NumQuizzesAttempted |
| PercentQuizzesAttempted | MaxScoreTotal |
| FirstScoreTotal | PercentMaxScore |
| PercentFirstScore | |
| **Forum Activities** | |
| ForumViews | ForumComments |
| ForumPosts | |
| **Summary Measures of Interaction** | |
| NumActiveDays | PercentDaysActive |
| DaysActiveWeek | |

Table 3: Features extracted and analyzed in this study.

### Data Analysis Procedure

Because our definition of engagement goes beyond predicting a binary outcome, we considered methods that could accommodate outcomes with three or more possible outcomes. In previous work, ordinal logistic regression has been used to predict persistence in MOOCs [10]. Because our definition of engagement is ordinal, in the sense that we consider more engagement to be better than some engagement or little engagement, we used an ordinal logistic regression model.

The specific ordinal logistic regression model we used is a form of the proportional odds model, which is defined by Agresti [1] as:

$$\text{logit}[P(Y \leq j)|\mathbf{x}] = \beta_{0j} + \beta'\mathbf{x}, \tag{1}$$

for $j = 1, 2, ..., J-1$. In our case, we have $J = 3$, corresponding to the different levels of engagement. Assuming the model maintains the proportional odds assumption, we have $J-1$ intercepts, but one coefficient for each independent variable; that is, the slopes are equivalent for each level of engagement,

[3] Jiang and colleagues [21] found that the week one quiz scores was predictive of receiving a certificate in a biology course to prepare for college; although the feature had a slightly different goal in their analyses, we considered it here.

but their intercepts are different. We used a *penalized* form of the proportional odds model; we investigated various levels of penalty, which shrink coefficients for unimportant independent variables towards, and eventually to, zero. Our goal was to reduce complexity and maximize interpretability to make our models actionable.

To build these models, we used the `ordinalNet` package in R [34] which estimates ordinal logistic regression models using the elastic net penalty. Models estimated using the elastic net penalty are well suited to situations when data are sparse [17], which is particularly accommodating to the data we have– where students might not be very active and might have very little information recorded for the variables in Table 3. When training each of these models, we needed to determine the quantity of the elastic net penalty, $\alpha$, which ranged between 0 and 1. When $\alpha = 0$, the technique is called the Ridge regression [19], and when $\alpha = 1$, the technique is called the least absolute shrinkage and selection operator (LASSO) [31]. After $\alpha$ was determined, we then needed to determine the optimal tuning parameter, $\lambda$. To determine the optimal parameters $\alpha$ and $\lambda$, we analyzed each course separately and tuned the parameters using 10-fold cross validation.

In our analyses, we experimented with seven different values for the elastic net penalty. For each $\alpha$ value, $\lambda$ values were tuned to achieve the maximum average out-of-sample loglikelihood. For all courses except Elearning, the maximum average was obtained using $\alpha = 1$; for Elearning, $\alpha = 0.9$ had a slightly larger maximum average than when we used the LASSO penalty, but the average out-of-sample misclassification rate was smaller when using the LASSO penalty. Given that the LASSO has desirable properties of shrinking coefficients and selecting important variables [13], and that the results from our empirical investigations show that when $alpha = 1$, the average out-of-sample log-likelihood is maximized, we used the LASSO penalty to further refine $\lambda$.

After assuming $\alpha = 1$, we used 10-fold cross validation to determine the $\lambda$ that maximized the average out-of-sample log-likelihood. We tested 50 $\lambda$s, starting with the harshest $\lambda$ for each MOOC model, which shrinks all but one coefficient (except the intercepts) to zero. The *lambda* values are sequentially smaller than the previous $\lambda$ values; thus, as $\lambda$ gets smaller, there are more nonzero coefficients. We used the one-standard-error rule [17] to determine the $\lambda$ used in our final model. That is, we used a larger $\lambda$ than the one that maximizes the average out-of-sample log-likelihood.

Once we determined reasonable values for the elastic net penalty $\alpha$ and the tuning parameter $\lambda$, we fit models for each of the courses. The `ordinalNet` package uses a coordinate descent algorithm to obtain parameter estimates for the penalized ordinal logistic regression models. When training these models, we did not penalize the intercept term. The penalized methods do not require the independent variables in the regression model to be independent; thus, we did not exclude levels of categorical variables, so there are no reference categories. This resulted in 71 independent variables for all courses except Elearning, which had 65 independent variables (because this course did not have any quizzes).

## RESULTS

We estimated five models of engagement, one for each course. Table 4 gives an overview of the features from Table 3 that shared a relationship with engagement for these courses. As a result of using the LASSO penalty, many coefficients for the 71 (or 65 for Elearning) predictors were shrunk to zero. The number of non-zero coefficients varied across the models: the Android model had 22 coefficients; Ochem and Elearning models each had 12 coefficients; the Subsistence model had 19 coefficients; and the Sustainability model had 16 coefficients. Note that because we used proportional odds models, we have two intercepts, corresponding to $P(engagement \leq 1)$ and $P(engagement \leq 2)$, but the coefficients for each independent variable were the same for each level of engagement [1]. When interpreting each independent variable, we assumed all others were held constant. To explore these patterns in more detail, we examined how the independent variables related to engagement for each of the five courses.

### Android Course Results

The model for the Android course revealed that students who were active many days each week (DaysActiveWeek) were more likely to be engaged for fewer than 3 weeks or 6 weeks. Similarly, increases in viewing lectures, the length and percentage of lectures viewed, the number and percentage of quizzes attempted, and the number of days active overall all had negative $\log(odds)$, suggesting that the more students engaged in these activities, the less likely they were to be classified in one of the higher levels of engagement. On the other hand, having social interaction–viewing (weeks 2 and 3), commenting (weeks 1 and 2), and posting on the forums (week 3)–during certain weeks was related to being engaged for at least 3 weeks or at least 6 weeks.

### Ochem Course Results

For Ochem, the model had negative coefficients for the following independent variables: AnyLectureActivity, TotalDownloadOnly, TotalLecturesWatched, PercentLecturesWatched, NumQuizzesAttempted, and PercentQuizzesAttempted. Thus, increases in lecture activity and quiz activity related to lower levels of engagement. Specifically, for an increase in any of these activities by a student, the lower that student's odds of being engaged for at least 3 or 6 weeks. Additionally, the more days a student was active during weeks 3, 4, 6, or 7, or an increase in either the number or percent of days active, resulted in that student's odds increasing for being engaged less than 6 weeks or less than 3 weeks.

### Subsistence Course Results

The indicator for earning an SOA was negative, suggesting that a student who earned an SOA was less likely to be engaged longer. Students in different sections had differing odds of being engaged longer or shorter, while students in North America were more likely to be engaged less than 6 weeks or less than 3 weeks. We found similar effects for AnyLectureActivity, PercentLecturesWatched, NumQuizzesAttempted, and PercentQuizzesAttempted: increases in these activities indicated that students tended to not be engaged for at least 3 or 6 weeks. As with the other courses, the more days active

| Feature | Course | | | | |
|---|---|---|---|---|---|
| | Android | Ochem | Subsistence | Sustainability | Elearning |
| Section | $\pm$ | | $\pm$ | $-$ | |
| Earned an SOA | | | $-$ | | |
| Age Group | | | | $-$ | |
| Timezone | | | $-$ | $-$ | |
| AnyLectureActivity | $-$ | $-$ | $-$ | $-$ | $-$ |
| TotalDownloadOnly | $-$ | $-$ | | | |
| TotalLecturesWatched | | $-$ | | | $-$ |
| LengthLecturesWatched | $-$ | | | | |
| PercentLecturesWatched | | $-$ | $-$ | $-$ | |
| PercentLengthWatched | $-$ | | | | $-$ |
| NumQuizzesAttempted | $-$ | $-$ | $-$ | $-$ | |
| PercentQuizzesAttempted | $-$ | $-$ | $-$ | | |
| ForumViews* | $+$ | | $+$ | | $+$ |
| ForumComments* | $+$ | | $+$ | | |
| ForumPosts* | $+$ | | | $+$ | |
| NumActiveDays | $-$ | $-$ | $-$ | $-$ | $-$ |
| PercentDaysActive | | $-$ | $-$ | $-$ | $-$ |
| DaysActiveWeek* | $-$ | $-$ | $-$ | $-$ | $-$ |

Table 4: Summary of LASSO models for engagement. $+$ denotes a positive relationship with engagement, $-$ denotes a negative relationship with engagement, and $\pm$ denotes some predictors have a positive relationship with engagement and others have a negative relationship. Entries omitted have regression coefficients equal to zero. For features that were recorded weekly (denoted by *), at least one coefficient for one week of the course was either positive or negative.

each week (except the first week of the course) were related to students not being engaged for at least 3 or at least 6 weeks. Similarly, increases in the number and percentage of active days resulted in an increase in odds of students being classified as having lower levels of engagement. Importantly, we found social interaction to be positively related to engagement: students who viewed the forum during the first week and those who commented in the forum during the second week had larger odds of being engaged in the course longer than those who did not.

#### Sustainability Course Results

The model for the Sustainability course revealed that increases in AnyLectureActivity, PercentLecturesWatched, and NumQuizzesAttempted resulted in larger odds of being in a lower level of engagement. Increases in the number of days active in some weeks of the course, as well as the NumActiveDays and PercentDaysActive resulted in smaller odds of being engaged over the weeks of the course. Hence, students who explored lectures, took quizzes, and were active many days over the weeks of the course had larger odds of being engaged less than 6 or less than 3 weeks. Students in one of the sections, as well as students in Africa, had larger odds of being engaged fewer weeks. Older students (i.e., those 50 and older) were less likely to be engaged over the weeks of the course. As with the Android and Subsistence courses, we found social engagement was related to longer engagement: students posting in the forum during the first week of the course were more likely to be engaged at least 3 weeks.

#### Elearning Course Results

For the Elearning course, AnyLectureActivity, TotalLecturesWatched, and PercentLecturesWatched all had negative $\log(odds)$; hence, increases in each of these activities meant that students were less likely to be engaged longer. For NumActiveDays and PercentDaysActive, we found that the more days a student was active, the more likely that student was to be engaged for fewer weeks. Furthermore, the more days per week a student was active in weeks 3, 4, 5, 6, 7, and 8, the larger that student's odds of being active fewer weeks. Students consuming lecture content and being active many days over the course were more likely to be engaged fewer weeks. However, for each view of the forum during week one, a student was more likely to be engaged at least 3 weeks.

#### Result Summary

As is evident from the previous description of each model, and Table 4, several patterns can be observed. Three of the independent variables, AnyLectureActivity, NumActiveDays and DaysActiveWeek, were associated with engagement across all five of the courses. Thus, students who participated in the lectures in any possible manner (i.e, streaming or downloading lecture videos) had higher odds of being engaged in one of the lower two engagement categories. Additionally, increases in the number of days active, over the entire course, and during each week, resulted in higher odds of being engaged fewer weeks. Additionally, NumQuizzesAttempted was negatively related to engagement in the four MOOCs that had quizzes; hence, holding all else constant, for each quiz a student took in these MOOCs, their odds of being engaged decreased.

Looking at the variables positively associated with engagement, we found that forum activities during some weeks shared a positive relationship with engagement in all courses except for Ochem: students who viewed the forums in Android, Subsistence, and Elearning had positive odds of being engaged more weeks. Posting comments in response to others' posts on the forums was associated with higher levels of engagement in Android and Subsistence, while posting original comments in the forums was associated with higher levels of engagement in Android and Sustainability. As a result, we found some types of social involvement in these MOOCs, by way of the forums, was associated with being engaged longer in the course.

## DISCUSSION

Our findings speak to other investigations examining if or when a student might drop out of a MOOC. Whereas many of the studies in Table 1 considered a single course, we considered five MOOCs, from different disciplines, offered over multiple iterations. Of the studies we referenced in Table 1, only six explored multiple courses. Furthermore, the set of features (in Table 3) that we used to predict engagement included a host of prediction channels versus just a few facets of the course. However, given that each of these studies investigated behaviors that are indicative of success (with various definitions) in MOOCs, in this section we compare our main findings with previous findings. Finally, we also discuss implications for MOOC design and opportunities for future work.

### Main Findings

We found that students who were active in the forums (by posting original comments, commenting on others' postings, or viewing the forum postings) had higher odds of being engaged longer in four of the courses we considered. To this end, participation in forums, or features derived from forum participation, have been identified in many studies as being related to completion or dropout in MOOCs, regardless of how success is defined (e.g., [8, 10, 21, 26, 27, 32, 35]). In general, many of these studies found that active participation in forums (the primary means of collaboration between students) was related to success (e.g., [8]). In fact, in the predictive models of Ramesh and others [27], predictors for forum participation were second only to lecture activity for predicting dropout. Given our finding that forum activities were positively related to engagement, we hypothesize that having a social connection to others in the course is related to being engaged for more weeks.

Surprisingly, we found that lecture activity and quiz-taking activities shared a negative relationship with engagement. Although it has been documented that when there are long videos to watch each week, students tend to skip them [16], we did not expect that consuming the lectures would be related to less engagement. This finding suggests that students engage in MOOCs for a variety of reasons, and may disengage with MOOCs both for similar and for entirely different reasons from the reasons they signed up for the MOOC in the first place. Perhaps the students in these MOOCs were bored by the videos or were only interested in just a small portion of the videos in these MOOCs because it has been documented that a significant proportion of students in MOOCs might be there just to audit or browse [28].

Additionally, we found the number of days active, both for the entire course, and within each of the eight weeks, shared a negative relationship with engagement. Being active many days in a week may relate to interest in the content for that week, but not for other weeks; hence a students' longevity, or lack thereof, may be predicted by students having accessed the content they wanted to access. Alternatively, being active many days per week may indicate students who require more time to learn the material, but had not planned in advance to spend more time. This would align with the findings of Greene et al. [14], who found that the number of hours a student planned to commit to a MOOC was related to longevity in the course. Furthermore, being active for many days each week might be indicative of poor time management. In previous studies, students who were found to be on-track [25] were more likely to be successful, and having a "balanced behavior pattern over the course of a week" [24] was an important predictor of drop out. Hence, with improved time management, students might be more engaged.

### Implications for MOOC Design

The findings in this paper can inform stakeholders in these and similar MOOCs, including universities, instructors, and course designers. For example, some students use MOOCs to test the waters before enrolling in a particular university [20]; thus, it may be beneficial to keep students engaged in courses as long as possible, because it may positively influence their perceptions when making future choices about which universities to apply to or attend. Given that forum participation was positively related to engagement in most of the MOOCs we considered, encouraging students to be active participants in the forums might lead to benefits, particularly as forum participation provides students with opportunities to become socially engaged with the course and by interacting with fellow students, and lead them to stay longer.

### Limitations and Future Work

One limitation of this investigation was the lack of control conditions in the courses evaluated. Control conditions examining MOOCs that do not have discussion forums (or mandatory discussion forums), for example, may illuminate reasons that findings differed across courses. We examined five different MOOCs and, for each of them except Ochem, increased forum activity was related to higher odds of engagement in the course. It is unclear from these data whether forum activity actually drives student engagement or whether students who were already more engaged simply preferred forum activities to other activities. However, previous investigations have suggested that forum participation is a powerful enticement for engagement in the course, and may be a powerful tool for learning [9]. Future investigations with content similar to that in the Ochem MOOC we investigated here might include attempts to restructure forum use, thereby seeing whether students can find value in Ochem forums or whether forums may be a distracting or a non-useful component of a MOOC devoted to the specific content of this MOOC, or other, related content. In sum, we note that this investigation adds nuance to others' findings that participation in forums is related to improving students' course performance [6, 12] by finding that participation in the forums is related to the level of longevity in the course and by adding the question of why some courses (here, Ochem) might be different from other courses.

In addition, because of the relatively consistent and positive findings for forum participation in relation to student engagement with and success in a MOOC, we wonder how far this realization can be taken. Even if it was only students with particular goals who showed these positive forum participation patterns, would it be possible to encourage all students to participate in forums, independent of their reasons for taking a MOOC? If so, would we obtain comparable results? In other words, are the benefits that come with active engagement in course forums enough of a hook or enough of a satisfying learning experience to overwhelm even those who come to a MOOC to dabble or learn a single piece of content offered in that MOOC? We hope to answer this question with an experiment to test the impact of how the course supports forum engagement in a future investigation.

## CONCLUSION

Understanding how students engage with MOOCs is crucial for effective development of course materials, deciding what functionalities should be included in a course, and evaluating students beyond grades or course completion. Toward this end, in this paper we described a measure of student engagement in MOOCs, which offered advantages over past measures that focus on dropout, and found that several patterns of student engagement were consistently present across five diverse course topics. To do this, we used the LASSO penalty for ordinal logistic regression models, which allowed us to ascertain which features were related to this definition of engagement. We found, across five MOOCs, that actively pursuing lectures, taking quizzes, and the number of days active overall and each week were related to being engaged fewer weeks. On the other hand, we found that students who were involved in the forums had higher odds of being engaged more weeks in all of the courses except one. These results are particularly important given recent research showing that findings in MOOCS often do not replicate across courses [3]. In the future, we will continue to explore how engagement can be characterized in MOOCs, and apply our findings to the design of future courses, thus moving towards providing students with a learning environment that leads to successful learning outcomes.

## ACKNOWLEDGEMENTS

We gratefully acknowledge the generous support of the Illinois Learning Sciences Data Initiative (ILSDI), at the University of Illinois Urbana-Champaign. The opinions expressed are those of the authors and do not represent views of the ILSDI.

## REFERENCES

1. Alan Agresti. 2002. *Categorical Data Analysis* (2nd. ed.). Wiley Interscience.
2. Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2014. Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web*. ACM, 687–698.
3. Juan Miguel L. Andres, Ryan S. Baker, George Siemens, Dragen Gašević, and Catherine A. Spann. 2017. Replicating 21 findings on student success in online learning. *Technology, Instruction, Cognition, and Learning* 10, 4 (2017), 313–333.
4. PG de Barba, Gregor E Kennedy, and MD Ainley. 2016. The role of students' motivation and participation in predicting performance in a MOOC. *Journal of Computer Assisted Learning* 32, 3 (2016), 218–231.
5. Christopher Brooks, Craig Thompson, and Stephanie Teasley. 2015. Who you are or what you do: Comparing the predictive power of demographics vs. activity patterns in massive open online courses (MOOCs). In *Proceedings of the Second (2015) ACM Conference on Learning@Scale*. ACM, 245–248.
6. Cho Kin Cheng, Dwayne E. Paré, Lisa-Marie Collimore, and Steve Joordens. 2011. Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance. *Computers & Education* 56, 1 (2011), 253–261.
7. Doug Clow. 2013. MOOCs and the funnel of participation. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. ACM, 185–189.
8. Scott Crossley, Mihai Dascalu, Danielle S. McNamara, Ryan S. Baker, and Stefan Trausan-Matu. 2017. Predicting success in massive open online courses (MOOCs) using cohesion network analysis. In *Making a Difference: Prioritizing Equity and Access in CSCL, 12th International Conference on Computer Supported Collaborative Learning (CSCL) 2017, Volume 1*. Philadelphia, PA: International Society of the Learning Sciences, 103–110.
9. Scott Crossley, Luc Paquette, Mihai Dascalu, Danielle S. McNamara, and Ryan S. Baker. 2016. Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, 6–14.
10. R. Wes Crues, Genevieve M. Henricks, Michelle Perry, Suma Bhat, Carolyn J. Anderson, Najmuddin Shaik, and Lawrence Angrave. 2018. How do gender, learning goals, and forum participation predict persistence in a computer science MOOC? *ACM Transactions on Computing Education* (2018).
11. Jennifer DeBoer, Andrew D. Ho, Glenda S. Stump, and Lori Breslow. 2014. Changing "course": Reconceptualizing educational variables for massive open online courses. *Educational Researcher* 43, 2 (2014), 74–84.

12. Lu Ding, ChanMin Kim, and Michael Orey. 2017. Studies of student engagement in gamified online discussions. *Computers & Education* 115 (2017), 126–142.

13. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2009. *The elements of statistical learning* (2nd ed.). Springer Series in Statistics.

14. Jeffrey A. Greene, Christopher A. Oswald, and Jeffrey Pomerantz. 2015. Predictors of retention and achievement in a massive open online course. *American Educational Research Journal* 52, 5 (2015), 925–955.

15. Philip J. Guo and Katharina Reinecke. 2014. Demographic differences in how students navigate through MOOCs. In *Proceedings of the first ACM conference on Learning@Scale*. ACM, 21–30.

16. Sherif Halawa, Daniel Greene, and John Mitchell. 2014. Dropout prediction in MOOCs using learner activity features. *Experiences and best practices in and around MOOCs* 7 (2014), 3–12.

17. Trevor Hastie, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical learning with sparsity: The lasso and generalizations*. CRC press.

18. Jiazhen He, James Bailey, Benjamin I. P. Rubinstein, and Rui Zhang. 2015. Identifying at-risk students in massive open online courses. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 1749–1755.

19. Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 1 (1970), 55–67.

20. Jason Paul Howarth, Steven D'Alessandro, Lester Johnson, and Lesley White. 2016. Learner motivation for MOOC registration and the role of MOOCs as a university 'taster'. *International Journal of Lifelong Education* 35, 1 (2016), 74–85.

21. Suhang Jiang, Adrienne Williams, Katerina Schenke, Mark Warschauer, and Diane O'dowd. 2014. Predicting MOOC performance with week 1 behavior. In *Educational Data Mining 2014*.

22. René F. Kizilcec and Sherif Halawa. 2015. Attrition and achievement gaps in online learning. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. ACM, 57–66.

23. René F. Kizilcec, Chris Piech, and Emily Schneider. 2013. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge*. ACM, 170–179.

24. Marius Kloft, Felix Stiehler, Zhilin Zheng, and Niels Pinkwart. 2014. Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*. 60–65.

25. Tommy Mullaney and Justin Reich. 2015. Staggered versus all-at-once content release in massive open online courses: Evaluating a natural experiment. In *Proceedings of the Second (2015) ACM Conference on Learning@Scale*. ACM, 185–194.

26. Jiezhong Qiu, Jie Tang, Tracy Xiao Liu, Jie Gong, Chenhui Zhang, Qian Zhang, and Yufei Xue. 2016. Modeling and predicting learning behavior in MOOCs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 93–102.

27. Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. 2014. Learning latent engagement patterns of students in online courses. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

28. Justin Reich. 2014. MOOC completion and retention in the context of student intent. *EDUCAUSE Review Online* (2014).

29. Carly Robinson, Michael Yeomans, Justin Reich, Chris Hulleman, and Hunter Gehlbach. 2016. Forecasting student achievement in MOOCs with natural language processing. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, 383–387.

30. Ayse Sunar, Su White, Nor Abdullah, and Hugh Davis. 2017. How learners' interactions sustain engagement: A MOOC case study. *IEEE Transactions on Learning Technologies* 10 (2017). Issue 4.

31. Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

32. Miaomiao Wen, Diyi Yang, and Carolyn Penstein Rosé. 2014. Linguistic Reflections of Student Engagement in Massive Open Online Courses.. In *ICWSM*.

33. Jacob Whitehill, Kiran Mohan, Daniel T. Seaton, Yigal Rosen, and Dustin Tingley. 2017. Delving Deeper into MOOC Student Dropout Prediction. *CoRR* abs/1702.06404 (2017). `http://arxiv.org/abs/1702.06404`

34. M. J. Wurm, P. J. Rathouz, and B. M. Hanlon. 2017. Regularized ordinal regression and the ordinalNet R package. *ArXiv e-prints* (June 2017).

35. Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. 2013. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*, Vol. 11. 14.

36. Saijing Zheng, Mary Beth Rosson, Patrick C. Shih, and John M. Carroll. 2015. Understanding student motivation, behaviors and perceptions in MOOCs. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM, 1882–1895.