Integrating Analytics and Surveys to Understand Fully Engaged Learners in a Highly-Technical STEM MOOC

Nathan M. Hicks[1], Doipayan Roy[2], Siddharth Shah[2], Kerrie A. Douglas[1], Peter Bermel[2], Heidi A. Diefes-Dux[1], and Krishna Madhavan[1]
[1] School of Engineering Education
[2] School of Electrical & Computer Engineering
Purdue University
West Lafayette, Indiana, USA

***Abstract*— Massive Open Online Courses (MOOCs) offer the ability to educate large numbers of diverse learners who might not have access, time, or the financial resources necessary for more formal coursework. While some studies have focused primarily on understanding MOOC learners purely through their access rates to course materials, others have sought to understand learners through surveys. We combined these two sources of data to address two research questions: (1) What are the patterns of user behavior in an advanced, technical MOOC? and (2) What are the characteristics of fully engaged learners? By analyzing clickstream and pre-survey data for a nanotechnology-related MOOC, we identified differences and similarities between fully engaged learners and other groups. The lack of strong indicators to predict fully engaged learners suggests a need for improved data from pre-course surveys.**

Keywords—Massive open online course; Learner analytics; Learner behavior; k-means clustering; Course design

I. Introduction

The low cost and highly available nature of massive open online courses (MOOCs) has generated significant appeal as a means to educate a diverse set of learners who might not have access to formal education. However, retention of MOOC learners proves challenging, likely due to low barriers to entry, variable learner intentions, and extreme flexibility of asynchronous instruction. As a result, some scholars argue that the educational community should reconceptualize the assessment of factors like retention in MOOCs [1]. While the majority of learners may treat MOOCs like textbooks—using course resources based on topical interest—some MOOC users participate in most or all aspects of the course [2]. Thus, the appropriateness of completion as a metric of course quality depends upon whether the MOOC developers or users value completion over achievement of individual learning goals.

Still, very little research has been conducted to characterize learners who fully engage in the majority of course materials, instead tending to focus on learners who drop out. The percentage of learners who fully engage may be small, but in many cases, the number of users who participate in most aspects of the MOOC greatly exceeds the number of students who would enroll in that course were it only offered at one university. Therefore, the validity of retention as an index for MOOC evaluation is indeed nuanced. While the call has been made to update traditional metrics for educational evaluation, little guidance has been given to actually inform what the new set of metrics should be for open online learning environments.

This problem of determining effective means of evaluation for MOOCs has been exacerbated by their rise in popularity. Since 2008, when the University of Manitoba offered the first MOOC (titled Connectivism and Connected Knowledge), the number of MOOCs and the topics covered have expanded dramatically [3]. The content presented across the collection of all MOOCs appeals to a continually widening pool of potential users. However, reasons for participating in MOOCs vary by user; some take courses for academic programs or professional development while others take courses purely out of curiosity [4]. This variance in purpose of MOOC enrollment may be partly associated with a number of factors, including the age, gender, ethnicity, field of study, and the content area of the course [4]. It is reasonable to speculate that the reason for taking the course would influence intended use and, therefore, that a number of variables may serve as indicators or predictors for behavioral patterns. Thus, understanding these user groups and their expectations when enrolling in a MOOC represents an important goal for optimizing the design of MOOCs to better meet user needs.

This consideration of diverse user groups and their learning needs prompts a few questions regarding how MOOCs are designed and for what audience. Basic design principles start from first understanding intended users and their needs. To optimize MOOCs in terms of pedagogies used to support learning, it is essential to understand who the users are, rather than considering all learners to be in the same group. Certainly, when examining whole course means, MOOC users outside of the dominant group may be lost in the thousands of users, yet still constitute a substantial number of actual persons. This issue of diverse learners is particularly relevant in highly specialized advanced MOOCs.

The purpose of this paper is to better understand the most heavily engaged group of learners. Thus, we seek to answer two research questions: (1) What are the patterns of user behavior in an advanced, technical MOOC? and (2) What are the characteristics of fully engaged learners?

II. Literature Review

Foundational to instructional design is understanding learners [5]. Yet, while MOOC literature is in abundance, few researchers have begun to tackle the problem of a theoretically justifiable way to understand student behavior, performance, or motivations [6]. Whether survey data, assessment performance, or user access to MOOC materials, simply calculating averages or using statistics based on means of the entire population simply does not capture the complexity of specific learner groups.

A major hurdle to understanding learner needs is the detemination of whether there is a constant typology of MOOC learners across platforms and course content [2]. Kizilcec and colleagues examined clusters of user access to course materials in three computer science MOOCs at introductory, upper undergraduate, and graduate levels [7]. The researchers found four groups of learner use patterns: *Completers*, *Auditors*, *Disengagers*, and *Samplers*. Douglas et al. examined patterns of user behavior in an advanced nanotechnology course, offered by nanoHUB-U [2]. Enrollment in the course was considerably lower than typical MOOCs as it was offered on a technical website rather than through a major MOOC platform. The results of the study found five groups of learners: *Fully Engaged*, *Consistent Viewers*, *One-Week Engaged*, *Two-Week Engaged*, and *Sporadic Users*. *Fully Engaged* learners accounted for 17% of the overall course enrollment—much higher than the typically reported MOOC 'completion' rate of 5% [8]. In addition, Douglas et al. found significant differences in assessment performance between user groups, shedding light on potential reasons behind learners dropping the course after fully participating for one or two-weeks [2]. There is a need to continue to examine the patterns of user behavior across MOOCs and platforms to determine whether there is a generalizeable typology of MOOC learners, based on behavior, assessment performance, and other personal characteristics.

III. Background

NanoHUB-U was developed as an extension of a leading center for computational nanotechnology [9]. The instructors of courses offered through nanoHUB-U are researchers in cutting edge fields of nanotechnology. In many cases, the course topics offered are so new that no textbooks cover all the relevant material [2]. Designed for engineering and science professionals, the topics are highly advanced and not readily understandable by the general masses. Indeed, the vast majority of people around the world would not have the background needed to understand the content covered in these courses, let alone the desire to complete all aspects of a course. Yet, for those whose work or academic pursuits are related with the fast-growing fields related to nanotechnology, the information provided may not be available elsewhere. In this sense, nanoHUB-U offers access to content material that is otherwise only in academic journals. Furthermore, discussions with nanoHUB-U faculty revealed that their primary concern was for the few learners who take the opportunity to utilize all course materials offered in the MOOC. Akin to a celebrity catering performances to her most devoted fans, the faculty valued the small percentage of learners who fully engaged in all aspects of the course and wanted to meet the needs of those users.

With a very clear user group (scientists and engineers whose work relates to nanotechnology) in mind, nanoHUB-U developed the course "Fundamentals of Nanoelectronics: Basic Concepts" and offered it on the edX platform. The live version of the course consisted of four units spanning eight weeks. The course had 9,888 total enrollees.

IV. Research Methods

To help MOOC instructors meet the needs of their most engaged students, it is necessary to learn who those users are—to understand the people behind the clicks. Before this can be achieved, however, we must analyze the usage data from an entire, completed course in order to accurately describe the behavior patterns of the most engaged group and properly identify who fits within that category. Thus, we selected one fully completed and heavily enrolled MOOC, Fundamentals of Nanoelectronics: Basic Concepts. Through collected clickstream data, we clustered the users into discernible groups. We then matched the users within each cluster to their pre-survey responses to statistically describe those learners' backgrounds, goals, and intentions. Finally, we rejected any partially completed surveys. This allowed for comparisons of responses with members of the other clusters to investigate what differentiates the most engaged learners.

A. Behavioral Clustering

Cluster analysis refers to methods of grouping data in such a way that the grouped elements exhibit the greatest similarity [10]. In case of analyzing the data for learners of a MOOC, this refers to the similarity of their online access patterns, as determined by the clickstream data. For cluster analysis of this particular course, we used a k-means algorithm, which groups the data into 'k' partitions by minimizing the distance of individual data points from their cluster centroids [11]. The number of k values is chosen by an elbow test, which plots the variance associated with each value of k [11]. We then choose the value where the addition of more clusters does not significantly increase the explanation of the grouped data. The clickstream data for this study consisted of 196,836 total records or clicking incidents, but a subset of 36,000 records were used for clustering. This subset consisted of all records of accesses to "sequential" or "chapter" modules, as those modules represent the only possible way to access "video" or "problem" modules. EdX provided clickstream data for 2,756 users and each user had the potential to access any of the 53 modules in the course.

B. Cluster Descrption

To better understand our most engaged learners as people, we needed to connect our clickstream data to our richest source of personal data—our pre-surveys, which we previously analyzed for internal correlations without the inclusion of behavioral data [12]. Running the clickstream data through the clustering algorithm provided the set of edX user IDs representing members of each of the identified clusters. Through an association table, we matched the user IDs with the hash IDs generated by each pre-survey instantiation. This matching enabled the sorting of pre-survey data into the appropriate behavioral clusters. However, as is often the case, not all users participated in the non-mandatory surveys, and curiously, some surveys were started or completed by individuals without user IDs or usage data. Thus, our descriptive analysis initially investigated the size of each cluster and the proportion of users within each cluster who began and completed the surveys.

Next, we summarized sets of pre-survey responses to address three questions regarding the most engaged learners: (1) What are their backgrounds? (2) What brought them to the course? and (3) What are their behavioral intentions? Using basic descriptive statistics, we found the cluster's overall distributions of pre-survey responses corresponding to each question.

C. Cluster Comparisons

Knowing the overall distributions of our most engaged learners is helpful, but of limited usefulness without knowing whether the distributions of those variables differ from other clusters of learners. To investigate this, we performed a Chi-square test of independence for each pre-survey question, sorting respondents by cluster. The Chi-square test for independence is a non-parametric statistical test for nominal or ordinal data used to determine the likelihood that two variables are independent from one another [13].

As with any statistical test, proper use requires the data to meet a set of assumptions: the data are measured in frequencies; the categories of each variable are mutually exclusive; each subject can contribute to only one cell; the study groups are independent; the variables are categorical at either the nominal or ordinal level; and, at least 80% of the calculated expected values must be at least 5 [13]. As each participant is only included in one cluster and can only provide single responses, and as our answer choices are either nominal or ordinal, our frequency data meet each assumption except for the final assumption. The expected values assumption was met for all but three items: in two of those cases, the responses could be reasonably grouped to meet the assumption and the grouping is reported in the Results section; in the third case, this grouping was not reasonable, but results are still reported with that limitation in mind. Additionally, Chi-square tests should be used on a randomly selected sample, so we must acknowledge that our sample was limited by those who completed the pre-survey. Thus, our results may suffer from self-selection bias.

The Chi-square test of independence tests against the null hypothesis that the proportion of respondents within a cluster providing a given response to a question matches the proportion of all respondents providing the same response. The test produces a Chi-square value based on the degrees of freedom (the product of one less than the number of rows and one less than the number of columns of the contingency table), which can be interpreted to find the probability of occurrence for that specific contingency table, or p-values [13]. When the p-values were less than 0.05 (suggesting that a contingency table as extreme as the one tested is likely to occur by chance less than 5% of the time under the null hypothesis), we analyzed the individual cells of the most engaged learners using individual cell Chi-square values, accepting values of less than 1 as indicating negligible difference between observed values and expected values [13]. For Chi-square tests, effect size or statistical strength is captured by Cramer's V [13]. Cohen provides a scale to interpret Cramer's V [14].

Average intrinsic and extrinsic motivation scores were also compared across clusters using a Kruskal-Wallis test. This test is similar to one-way analysis of variance (ANOVA), as it compares means of an ordinal variable across nominal categories with the null hypothesis of equality. However, as our clusters did not have normal distributions, but had similar population variances, our data met the assumptions for Kruskal-Wallis [15].

V. RESULTS

A. Clusters Identified

An elbow plot is used to optimize the number of clusters selected. The elbow plot created by the k-means algorithm for our data set is shown in Fig. 1. The *y*-axis on the plot represents the ratio between the cluster sum of squares and the total sum of squares (and is therefore always less than or equal to 1). The *x*-axis is the corresponding number of clusters for which the ratio is computed. Generally, one should select the last noticeable elbow after which negligible gains are made for including additional clusters. Our plot has potentially selectable elbows at four, five, and seven clusters. However, to be consistent with our previous research on a different nanotechnology-related course, we selected five clusters [2].

Once clustered, the users were sorted and plotted based on "sequential" and "chapter" module accesses, as shown in Fig. 2. This plot strongly resembles the plot our group generated in the previously mentioned study for another course [2], so we adopted the same names for the clusters. From top to bottom, these clusters contain the following: fully engaged learners, consistent learners, two-week engaged learners, one-week engaged learners, and sporadic learners. These clusters consisted of 297 (10.8%), 182 (6.6%), 299 (10.8%), 543 (19.7%), and 1,435 (52.1%) users, respectively.

It is worth noting that the k-means clustering algorithm itself does not cluster based on which modules were accessed, but rather purely based on the number of modules accessed. We assumed that because the course follows a logical progression, learners with similar numbers of access points would have accessed similar modules. Fig. 2 indicates that this assumption is reasonable, as most users in a given cluster had similar access patterns. It also bears mention that the vertical lines correspond to end-of-unit exams, which appear to continue to attract members across all clusters, regardless of engagement throughout the rest of the course.

B. Survey Completions

While there were 2,756 users clustered, 1,451 users initiated the pre-survey. However, only 969 users completed the survey; the incomplete surveys were removed from the data set. Additionally, 218 of the survey completers were not in the clustered set of users and, thus, had to be removed from the data set. Of the remaining 751 users, 172 were "fully engaged", 102 were "consistent", 121 were "two-week" engaged users, 196 were "one-week" engaged users, and 160 were "sporadic" users. Fig. 3 shows a comparison of the proportions of users in each cluster who either completed, started but did not complete, or did not attempt the pre-survey. A Chi-square test shows the differences were significant ($X^2(8, N = 751) = 602.37$, $p < 0.001$, Cramer's V = 0.33). This result empirically demonstrates

Fig. 1. The elbow plot indicating the strength of behavioral representation based on the number of clusters

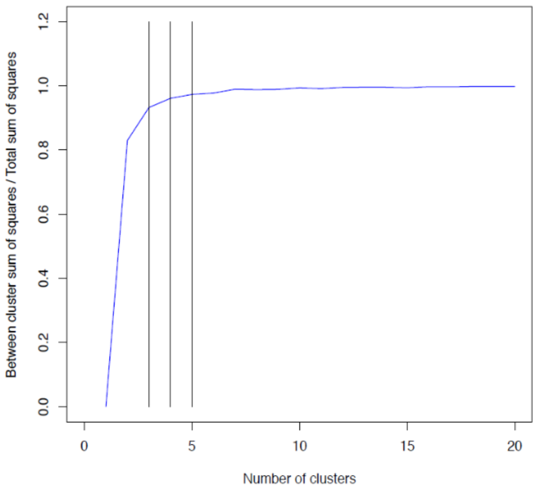
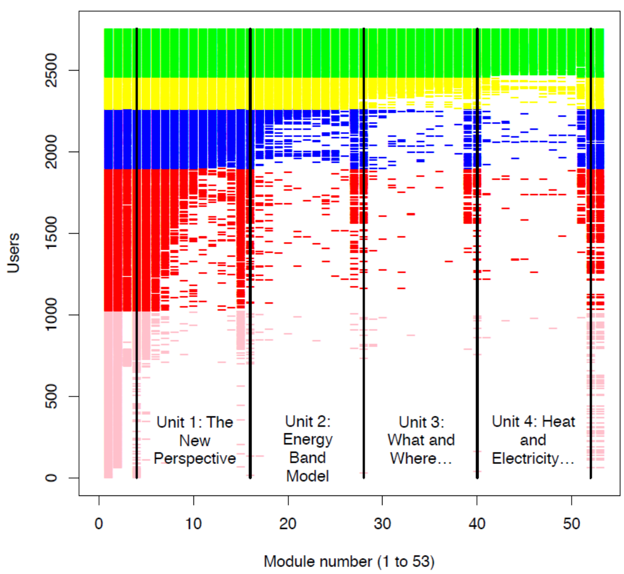


Fig. 2. Learner clustering by course material usage pattern



that those who are most engaged in the course are also most likely to participate in the course surveys—a finding which must be accounted for when analyzing future survey results.

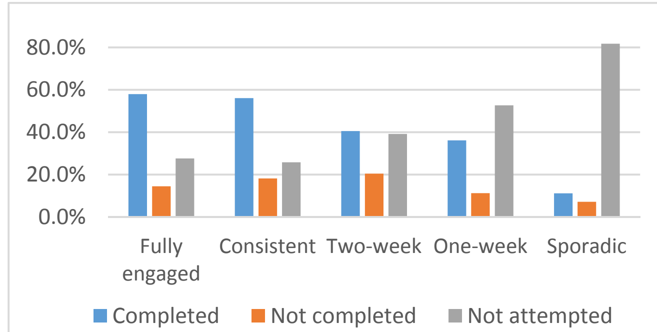
C. Fully Engaged Learner Statistics

The pre-survey items were sorted to answer three overarching questions: (1) Who are the fully engaged learners? (2) What are the conditions associated with fully engaged learners taking this course? and (3) What are the fully engaged learners' intentions and goals in taking this course?

(1) Who are the fully engaged learners? To answer the first question, we analyzed the responses to the following items: Is English your native language?

- What is your level of education?
- Have you had prior courses in this area of study?
- How much calculus have you taken?

Fig. 3. Proportional comparison of pre-survey completions by cluster



- Have you taken differential equations? and,
- What is your employment status?

The frequencies and percentages of "fully engaged" learners who provided each possible response to these items are reported in Table I. From this data, we see that the vast majority of "fully engaged" learners are not native English speakers, most have at least a bachelor's degree, most have taken prior courses related to the subject, most have taken greater than two semesters of college calculus, and most have taken differential equations. Many of these learners are also either employed full time, taking the course for professional development, or are full-time students.

Additional information regarding who these learners are can be obtained from their motivation scores. The pre-survey contained eight 7-point Likert-scale (strongly disagree to strongly agree) statements associated with motivation—four for intrinsic motivation and four for extrinsic motivation. The respondents' intrinsic scores were based on the average of their responses to intrinsic items and extrinsic scores were based on the average of their responses to extrinsic items. The mean of the "fully engaged" learners' intrinsic scores was 5.84 out of 7 with a standard deviation of 1.00, indicated that these learners were, on average, strongly intrinsically motivated. Their mean extrinsic score was 4.24 out of 7 with a standard deviation of 1.41, indicated nearly neutral extrinsic motivation. The distributions are shown in Figures 4 and 5, respectively.

(2) What are the conditions associated with fully engaged learners taking this course? The pre-survey had three items that explore this question:

- How did you first hear about nanoelectronics?
- Which best describes your reason for taking this course?
- Which best describes what you hope to gain from this course?

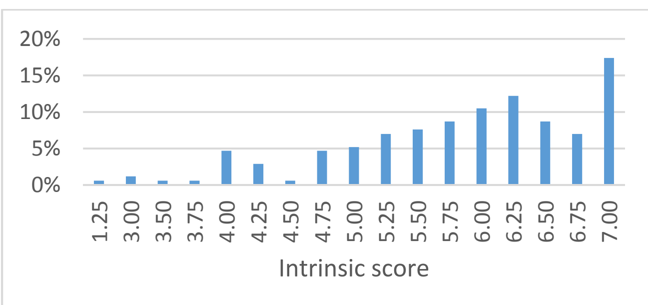
Table II presents the frequency and percentages of responses to each of these pre-survey items for the fully engaged cluster.

TABLE I. PRE-SURVEY RESPONSES TO ITEMS THAT ADDRESS THE QUESTION, "WHO ARE THE FULLY ENGAGED LEARNERS?"

Question	Answer choice	Frequency	Percent
Is English your native language?	Yes	33	19.2%
	No	139	80.8%
What is your level of education?	Precollege	1	0.6%
	Some college	12	7.0%
	Associate's degree	2	1.2%
	Bachelor's degree	53	30.8%
	Master's degree	58	33.7%
	Ph.D.	46	26.7%
Have you had prior courses in this area?	Yes	131	76.2%
	No	41	23.8%
How much calculus have you taken?	None	2	1.2%
	High school level	8	4.7%
	1 semester of college	11	6.4%
	2 semesters of college	37	21.5%
	>2 semesters of college	114	66.3%
Have you taken differential equations?	Yes	147	85.5%
	No	25	14.5%
What is your employment status?[a]	FT work, PD	62	36.0%
	FT work, PT student	21	12.2%
	FT student	55	32.0%
	Not working, not student	34	19.8%

[a] FT = full-time; PT = part-time; and PD = professional development

Fig. 4. Distribution of intrinsic scores for "fully engaged" learners



The statistics show that the fully engaged learners were most likely to hear about nanoelectronics through edX, followed by their own personal study, and then through nanoHUB. Most learners took the course because they were "very interested" in the subject, though many also reported that they were merely "curious" about the subject. These answer choices suggest similar intent (taking the course out of interest), occurring to differing degrees. One could argue that the difference in answering one versus the other might stem from a difference in familiarity with the subject. The final item shows that this group was about equally as likely to want to become well-acquainted with the material as they were to hope to apply it. Either way, they were looking for a deeper understanding than the more superficial learning suggested by the first answer choice.

Fig. 5. Distribution of extrinsic scores for "fully engaged" learners

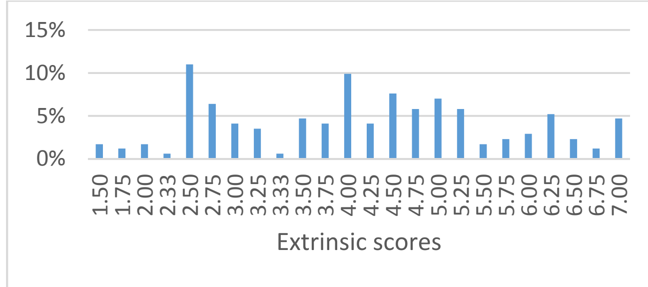


TABLE II. PRE-SURVEY RESPONSES TO ITEMS THAT ADDRESS THE CONDITIONS AROUND THESE LEARNERS TAKING THE COURSE

Question	Answer choice	Frequency	Percent
How did you first hear about nanoelectronics?	edX website	67	39.0%
	nanoHUB.org	26	15.1%
	Professional setting	16	9.3%
	Own personal study	31	18.0%
	Coursework	14	8.1%
	Other	18	10.5%
Which best describes your reason for taking this course?	Curious about nanoelectronics	53	30.8%
	Very interested	90	52.3%
	Course was recommended	14	8.1%
	Familiar with instructor	8	4.7%
	Other	7	4.1%
Which best describes what you hope to gain from this course?	A broad overview	38	22.1%
	Become well-acquainted	66	38.4%
	Be able to apply concepts	68	39.5%

(3) What are the fully engaged learners' intentions and goals in taking this course? The pre-survey had three items related to determining the learners' goals or intentions for the course:

- Which best describes your learning goals for the course?
- Which best describes your intended use of this course?
- How much time do you plan to dedicate each week to this course?

The responses to these questions are summarized in Table III. The results indicate that the fully engaged learners were somewhat divided between wanting a high grade in the course and not caring about course grades. The majority of these users anticipated participating in all aspects of the course. Further, the majority expected to dedicate between one and nine hours a week to the course, with the largest group expecting between three and six.

D. Comparisons between Clusters

While the data in the previous section give some indication of who the fully engaged learners are, why they are taking the course, and how they intend to use the course, it does not tell us whether any of those tendencies are unique for that cluster. To determine this, we performed a Chi-square test of independence for each of the pre-survey items presented in the previous section versus the five clusters. The results are presented in sub-sections corresponding to the sub-sections in which they were presented in the previous section.

TABLE III. PRE-SURVEY RESPONSES TO ITEMS ADDRESSING THE INTENTIONS AND GOALS OF THE FULLY ENGAGED LEARNERS IN THIS COURSE

Question	Answer choice	Frequency	Percent
Which best describes your learning goals for the course?	Achieve a high grade	77	45%
	Achieve a passing grade	35	20%
	Not concerned with grades	60	35%
Which best describes your intended use of this course?	Participate in all aspects	98	57.0%
	Watch all videos	22	12.8%
	Watch some videos	33	19.2%
	Some activities and videos	2	1.2%
	Decide after course starts	14	8.1%
	Just browsing	3	1.7%
How much time do you plan to dedicate each week to this course?	1–3 hours	31	18.0%
	3–6 hours	72	41.9%
	6–9 hours	33	19.2%
	9–12 hours	12	7.0%
	12–15 hours	7	4.1%
	Uncertain	17	9.9%

(1) Who are the learners?

a) Native language: We are able to reject the null hypothesis that frequencies of native English speakers are independent of cluster, $X^2(4, N = 751) = 10.64$, $p = 0.03$, Cramer's V = 0.12. The comparative proportions of users within each cluster who answered yes and no to this survey item are shown in Fig. 6. As the whole contingency table was found to be statistically significant, individual Chi-square values were calculated for each "fully engaged" cell, suggesting that significantly fewer fully engaged learners were native English speakers than was expected.

b) Education: The Chi-square test for cluster versus level of education resulted in failed assumptions, as more than 20% of the cells had expected values less than 5. However, when "Precollege," "Some college," and "Associate's degree" are joined into a "Less than bachelor's" category, the assumption is met and the results are significant, $X^2(12, N = 751) = 30.51$, $p = 0.002$, Cramer's V = 0.12 (percentages of answer by cluster are shown in Fig. 7). Individual cell Chi-square values suggest that this group is far less likely to have anything less than a bachelor's degree and far more likely to have a Ph.D.

c) Prior related coursework: Fig. 8 illustrates the percentage of learners who had prior related coursework, separated by cluster. The corresponding frequencies result in a statistically significant contingency table, $X^2(4, N = 751) = 9.832$, $p = 0.04$, Cramer's V = 0.11. Individual cell Chi-square values suggest that fully engaged learners were more likely to have previously taken a related course.

d) Calculus and Differential Equations: The Chi-square test of independence failed to reject the null hypotheses that the learners' experience with Calculus, $X^2(16, N = 751) = 25.38$, $p = 0.06$, Cramer's V = 0.09, and Differential Equations, X^2 (4, N = 751) = 9.09, $p = 0.06$, Cramer's V = 0.11, differs by cluster. It would seem that the majority of people willing to enroll in an advanced technical MOOC had similar mathematical skills, regardless of their ultimate participation.

Fig. 6. Native English speakers per cluster

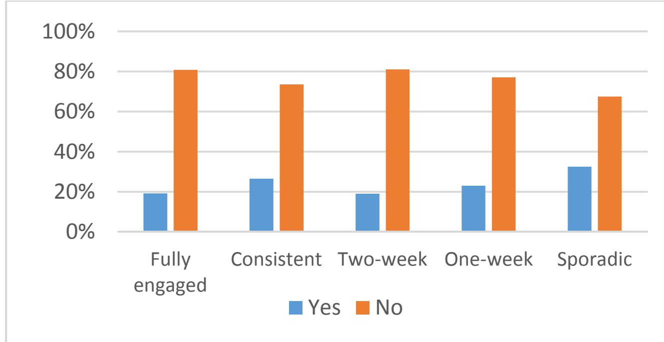
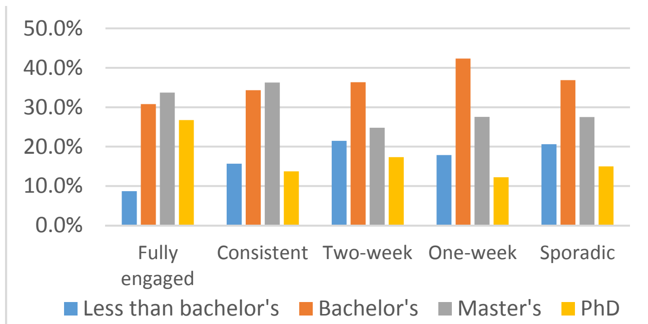
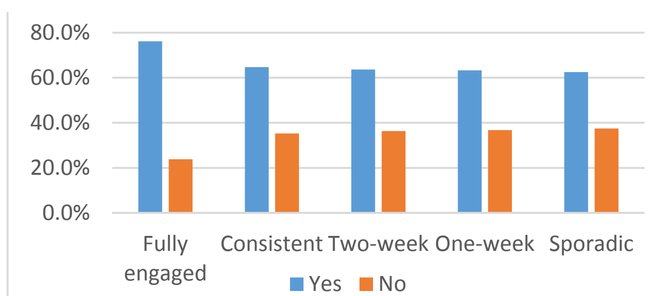


Fig. 7. Level of education per cluster



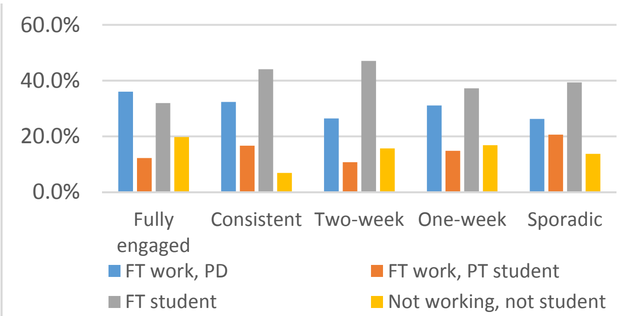
e) Work status: The pre-survey responses suggest that there is a statistically significant difference between clusters in terms of work status (see Fig. 9), $X^2(12, N = 751) = 21.97$, $p = 0.04$, Cramer's V = 0.10. Individual cell Chi-square values indicate fully engaged learners are more likely to be full-time employees seeking professional development or to be unemployed nonstudents and are less likely to be full-time students.

Fig. 8. Prior related coursework per cluster



f) Motivation: In addition to these variables, a Kruskal-Wallis test was conducted to compare means of the intrinsic and extrinsic motivation scores for each cluster. No significant difference existed for either: X^2 (4, N = 751) = 4.105, $p = 0.39$ and X^2 (4, N = 751) = 6.34, $p = 0.18$, respectively.

Fig. 9. Work status per cluster



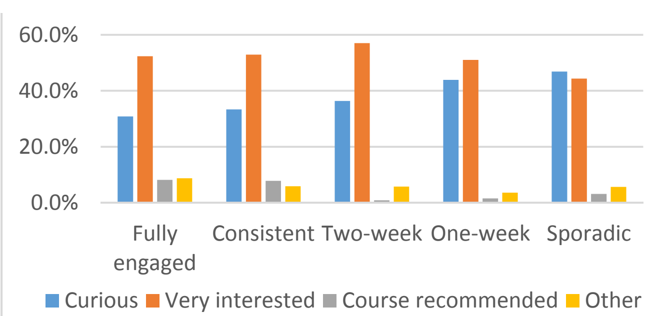
(2) What are the conditions associated with taking this course?

a) Where was concept heard? Overall, the contingency table for this item was not considered significant, X^2 (20, N = 751) = 28.65, p = 0.10, Cramer's V = 0.10. However, individual cell Chi-square values indicated that fully engaged learners were less likely to have heard the concept through edX and far more likely to have heard of the concept through nanoHUB.

b) Reason for enrolling: The contingency table for this item initially produced expected values less than 5 in 36% of the cells. Because this item included an "other" answer choice and because both "other" and being familiar with the instructor had such small numbers, those two answers were combined. The resulting contingency table met the required assumptions and yielded significant results, X^2 (12, N = 751) = 31.212, p = 0.002, Cramer's V = 0.12. Individual cell Chi-square values suggest that fully engaged users were less likely to take the course out of curiosity and more likely to take the course based on recommendation or other reasons. The modified results are summarized in Fig. 10.

c) Learning outcomes: The intended learning outcomes were similar across clusters, X^2 (8, N = 751) = 10.43, p = 0.24, Cramer's V = 0.08.

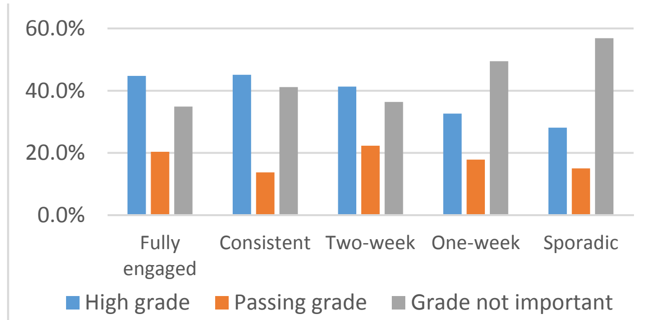
Fig. 10. Reasons for taking course across clusters ("Other" includes both survey options of "other" and being familiar with the instructor)



3) What are learners' intentions and goals?

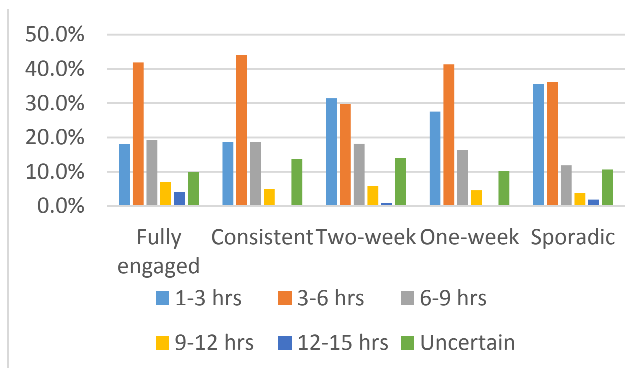
a) Grade sought: Learner responses, summarized comaratively in Fig. 11, exhibited a statistically significant difference across clusters, X^2(8, N = 751) = 25.29, p = 0.001, Cramer's V = 0.13. Individual cell Chi-square values indicate that fully engaged users were more likely to desire a high grade and less likely to say grades do not matter.

Fig. 11. Course grade sought across clusters



b) Time commitment: Fig. 12 summarizes the responses to this item. The Chi-square test shows that intended time commitment does vary significantly by cluster, X^2 (20, N = 751) = 37.53, p = 0.01, Cramer's V = 0.11. Pairwise z-tests indicate that the fully engaged learners were more likely to intend to spend 3-6, 6-9, 9-12, or 12-15 hours per week on the course than were other clusters (for the most part).

Fig. 12. Intended time commitment across clusters



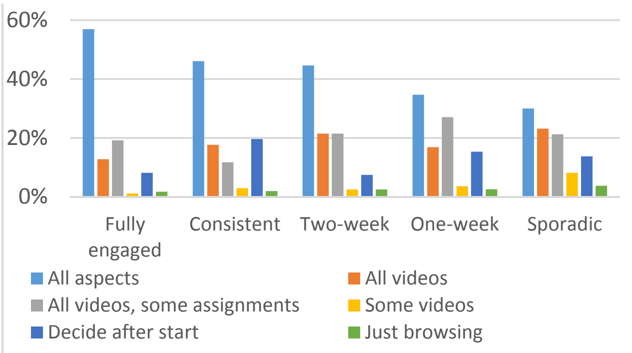
c) Intended Use: The contingency table (summarized in Fig. 13) for intended use of the course materials produced 23.3% of cells with expected values less than 5, barely failing to meet the assumptions for the Chi-square test. The survey item's answer choices did not seem to be reasonably combined, but the results of the test were significant, X^2 (20, N = 751) = 55.50, p < 0.001, Cramer's V = 0.14. Individual cell Chi-square values suggest that fully engaged learners were much more likely to intend to use all course materials. These results are reported acknowledging potential unreliability.

VI. Discussion

A. Fully Engaged Learners versus Other Clusters

Fully engaged learners were generally nonnative English speakers, had post-secondary degrees, had experience in the course subject, had strong mathematical backgrounds, and were more intrinsically than extrinsically motivated. They predominantly were full-time employees seeking professional development or were full-time students. Statistical tests of independence revealed that each of these descriptors of the

Fig. 13. Intended use of course matierals per cluster



cluster, except for the strong mathematics background and motivational tendencies, differed significantly from expected values based on the sample as a whole.

Fully engaged learners generally learned of nanoelectronics through edX, personal study, or nanoHUB, but were more likely to have heard of the subject through nanoHUB than other clusters. While the majority of fully engaged learners took the course out of curiosity or interest, they were more likely than other groups to take the course based on recommendation or other reasons, though the frequencies in these categories are small enough to warrant caution of interpretation. Also, most of these learners wanted to become at least well-acquainted with the course content, but these differences were not significant.

Fully engaged learners also generally wanted to earn a high grade, participate in all aspects of the course, and dedicate between three and six hours a week to the course. Each of these categories were shown to differ significantly from expected values based on the sample as a whole. As a result, this information can be used to design the class to more effectively target the fully engaged audience.

B. *Implications*

Knowing the basic common characteristics of the fully engaged learners may help instructors to cater to that specific audience. For instance, the fact that the majority of the fully engaged learners hold advanced degrees and have prior related coursework indicates that course content may be presented at an advanced level without significant concern for surpassing the learners' abilities. Further, knowing that many of the fully engaged learners intend to spend between three and six hours on the course per week, the course materials should be designed to require a little over three hours of work each week.

It should be noted, however, that, despite the predominant descriptors of fully engaged learners, none of those descriptors were unique for the fully engaged learner cluster. For instance, simply knowing that a learner has a Ph.D. does not guarantee that that user will be a fully engaged learner. Consequently, individual descriptors are not significant predictors of learner behavior. Further, our data do not tell us much about what the learners want out of the course, how much time they spend on the course, what factors outside of the course might motivate or demotivate them, or why their behaviors might not match their initial intentions. Thus, the pre-surveys must be revised to better understand learner intentions, goals, and motivations. Additional clickstream data should also be collected to determine the extent to which behavior matches intentions. And finally, post-surveys or interviews should be designed to capture why learners struggled to engage as intended.

C. *Future Work*

As none of the questions on the pre-survey seem to truly distinguish the fully engaged learners from the other clusters, it might be worth exploring other survey items to achieve greater differentiation. This would provide stronger information about the target audience. For instance, what is the ideal range of time needed to learn the concepts in this course for someone without significant experience in the area? The only direct connection between our clickstream data and the pre-survey questions is with intended use of course materials.

There are also a few select groups who might deserve further investigation. What happened to the users who initially indicated strong commitment to the course but ultimately were sporadic users? Did personal costs limit their available time? Or did they not believe they could succeed or did not see value in the course? Alternatively, what switched for learners who did not indicate much initial interest but then were fully engaged learners? These questions, coupled with the relative ineffectiveness of our intrinsic-extrinsic motivation measurements make a transition to an expectancy-value-cost motivational framework sound appealing as a future direction.

D. *Limitations*

This study provided a large number of statistical comparisons, increasing the likelihood of type I error (false positives). Additional studies of other courses and larger populations may be helpful to confirm general findings. This would also benefit the relatively small effect sizes found in this study. Further, the respondents were self-selected, which introduces bias. This is likely always a problem when looking at MOOC surveys—if survey completions were mandatory, we would likely suspect a large number of invalid responses. The completion behavior presented in the beginning of the results section supports this.

One final complication is that the k-means algorithm produces slightly different clusters upon each execution. In other words, one run may place user X in the "fully engaged" group, while a second run may place user X in the "consistent" group. Although some users will fluctuate across clusters in separate runs of the algorithm, the overall process is robust, so the majority of users will consistently be placed in the same cluster. Still, minor variance might be relevant when analyzing categories with small frequencies, which occurred for some survey items in our study.

VII. Conclusions

This study provides basic descriptive statistics for a heavily engaged group of learners in an advanced technical MOOC—a group often neglected in studies of MOOC users, as the focus is generally on dropouts. We provide general trends about this group and explore how they differ from other user groups in the hope of developing strategies to most effectively teach the group who invests the most commitment to their learning.

ACKNOWLEDGMENT

This work was made possible by a grant from the National Science Foundation (NSF DGE 1544259). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] J. DeBoer *et al.*, "Changing 'course': Reconceptualizing educational variables for massive open online courses," *Educ. Researcher,* vol. 42, no. 2, 74–84, 2014. doi:10.3102/0013189x14523038

[2] K. A. Douglas *et al.*, "Smart data characterization of a highly technical MOOC-like engineering course," submitted for publication.

[3] J. A. Greene *et al.*, "Predictors of retention and achievement in a massive open online course," *Amer. Educ. Research J.,* vol. 52, no. 5, 925–955, 2015. doi:10.3102/000831215584621

[4] G. Christensen *et al.*, "The MOOC phenomenon: Who takes massive open online courses and why?" unpublished.

[5] P. L. Smith and T. J. Ragan, *Instructional Design*. New York, NY: Wiley, 1999.

[6] E. Wiebe *et al.*, "MOOCs From the Viewpoint of the Learner A Response to Perna et al. (2014)," *Educ. Researcher*, vol. 44, no. 4, 252–254. http://doi.org/10.3102/0013189X15584774

[7] R. F. Kizilcec *et al.,* "Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses," In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, New York, NY, 2013, pp. 170–179. http://doi.org/10.1145/2460296.2460330

[8] K. Jordan, "Initial trends in enrolment and completion of massive open online courses," *Int. Review of Research in Open and Distance Learning*, vol. 15, no. 1, pp. 133–160, Feb. 2014.

[9] K. Madhavan *et al.*, "Learning and research in the cloud," *Nature Nanotechnology*, vol. 8, no. 11, pp. 786–789, 2013. doi:10.1038/nnano.2013.231

[10] S. Na *et al.*, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," in *Third International Symposium on Intelligent Information Technology and Security Informatics*, 2010, pp. 63–67. doi:10.1109/IITSI.2010.74

[11] R. Tibshirani *et al.,* "Estimating the number of clusters in a data set via the gap statistic," *J. R. Statist. Soc. B*, vol. 63, no. 2, pp. 411–423, 2001.

[12] N. Hicks *et al*., "Learners in advanced nanotechnology MOOCs: Understanding their intention and motivation," in *American Society for Engineering Education's 123rd Annual Conference & Exposition*, New Orleans, LA, 2016.

[13] M. L. McHugh, "The Chi-square test of independence," *Biochemia Medica*, vol. 23, no. 2, pp. 143–149, 2013. doi:10.11613/BM.2013.018

[14] J. Cohen, "Chi-square tests for goodness of fit and contingency tables," in *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988, pp. 215–272.

[15] T. Van Hecke, "Power study of anova versus Kruskal-Wallis test," *J. Stat. Manage. Syst.,* vol. 15, no. 2–3, pp. 241–247, 2012.