Who they are and what they want: Understanding the reasons for MOOC enrollment

R. Wes Crues
University of Illinois at Urbana-Champaign
crues2@illinois.edu

Nigel Bosch
University of Illinois at Urbana-Champaign
pnb@illinois.edu

Carolyn J. Anderson
University of Illinois at Urbana-Champaign
cja@illinois.edu

Michelle Perry
University of Illinois at Urbana-Champaign
mperry@illinois.edu

Suma Bhat
University of Illinois at Urbana-Champaign
spbhat2@illinois.edu

Najmuddin Shaik
University of Illinois at Urbana-Champaign
shaik@illinois.edu

ABSTRACT

The diversity in reasons that students have for enrolling in massive open online courses (MOOCs) is an often-overlooked aspect while modeling learners' behaviors in MOOCs. Using survey data from 11,202 students in five MOOCs spanning different academic disciplines, this study evaluates the reasons that students enrolled in MOOCs, using an unsupervised learning method, Latent Dirichlet Allocation (LDA). After fitting an LDA model, we used correspondence analysis to understand whether these reasons were general, and could be invoked across the five MOOCs, or whether the reasons were course-specific. Furthermore, log-linear models were employed to understand the relations between the reasons students enrolled, the course they took, and their background characteristics. We found that students enrolled for many different reasons, and that their age was statistically related to the reasons they gave for taking a MOOC, but their gender was not. The paper concludes with a discussion of how instructors and course designers can use this information when creating new—or redesigning existing—MOOCs.

Keywords

MOOCs, informal education, text mining

1. INTRODUCTION

Massive open online courses (MOOCs) have been celebrated because they offer education to wide groups of students who may not otherwise have access to their rich content; they provide access to well-respected experts; they have a relatively low cost; and they are convenient. On the other hand, MOOCs have been criticized because they have high attrition and low completion rates. We acknowledge that there are high attrition and low completion rates, but if students sign up with the intent of only learning some aspects of what is offered in the MOOC, and not necessarily with the intent to learn everything that the MOOC has to offer, this ought not to be considered a failure. Acknowledging that MOOC learners have different reasons for enrolling in MOOCs—for example, to improve their skills, gain access to new knowledge, or dabble in an area they find intriguing—we examine whether the reasons students offer are MOOC-specific or content-generic for five MOOCs. We do this with the intent to distinguish whether the reasons that learners have for enrolling in MOOCs is linked to their background (age or gender) or to the specific course they have enrolled in. By finding ways to classify these reasons reliably, we will be in position to understand the relation between why students enroll in these courses and how successfully they navigate the course.

Although it may be advantageous for students to participate in all aspects of a MOOC and to complete the course, MOOCs are beginning to accommodate different paths and different outcomes. For example, Coursera[1] (one of the most popular MOOC providers) offers verified course completion certificates for students who wish to obtain proof of their accomplishments, but also allows students to enroll for no credit and sample whatever course materials they wish. However, most MOOCs do little to support the multiplicity of learning objectives that students may have for taking a particular MOOC. Understanding students' reasons for enrolling in a MOOC could put instructors in the position to make accommodations, potentially improving students' learning experiences.

A growing body of literature has investigated why students enroll in MOOCs (e.g., [7, 3, 5, 16, 23, 27]). These studies have used survey methods with closed-form responses or have used interviews. With surveys using closed-form responses, students are forced to select from a list of reasons; and with interviews, typically, only a limited number of students may be reached. In the current study, we investigated more than 11,000 students enrolled in five MOOCs, across several disciplines, using Latent Dirichlet Allocation (LDA) [2] to analyze their responses to an open-ended survey. We then used the probabilities from the LDA model to assign

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

each student to one of the topics (i.e., reasons for enrolling) generated by the LDA model. After we found the most probable reason a student enrolled, we cross-classified students by their most probable topic (i.e., reason) and the course for which they enrolled. We then visualized these relationships using correspondence analysis.

In this paper, we contribute to understanding student behavior in MOOCs by examining the reasons that students offered for enrolling in MOOCs, and the extent to which these reasons are unique to the specific MOOCs or whether they apply more generally, across MOOCs. Additionally, we advance understanding by using LDA and the results of log-linear models to hone in on specific relationships between student background characteristics and their reasons for enrolling. Using these results, we conjecture about how instructors and course designers could use this information to improve their courses and their students' learning experiences, thus contributing to the discussion about improving instruction for diverse learners.

2. RELATED WORK

Several studies have sought to make sense of what kinds of students enroll in MOOCs, and why. Specifically, these studies have examined students' background characteristics and why they take MOOCs. We discuss some of these works in the following subsections.

2.1 Goals for Enrollment in MOOCs

Current findings on why students enroll in MOOCs have revealed that students enroll in these courses for many different reasons. Hew and Cheung [11] identified common trends for why students enrolled in MOOCs, including: (1) a general interest in learning; (2) a desire to receive formal recognition of their knowledge; (3) an intent to explore course content without a strong desire to receive such recognition; and, (4) an interest or general curiosity in taking a MOOC. Next, we explore some of these themes in more depth.

Zheng et al. [27] interviewed students who took MOOCs and asked about their reasons for enrolling in MOOCs. Some students in their study were fulfilling their current needs, such as supplementing a for-credit course, or to help with their current position, either as students or in a workplace setting. Other students offered that they took the course to develop a social connection with others who shared similar interests. Additionally, they found some who enrolled did so to prepare for future job opportunities or to gain experience in a field they might study in a more formal manner after taking the MOOC. Finally, some of the students in this investigation enrolled in the MOOCs because they were interested in satisfying (broadly) their curiosity. Along these lines, it has been posited that MOOCs function as previews of what might be offered to students in a for-credit university course [15].

Kizilcec and Schneider [16] developed the Online Learning Enrollment Intentions (OLEI) questionnaire, which asked students to select whether or not each of 13 different reasons for enrolling in a MOOC applied to them. These reasons included career-related interests, formal education, social opportunities, potential career benefits, personal enrichment, and prestige. Liu, Kang, and McKelroy [18] found most of the students in a set of MOOCs took those MOOCs for personal interest, or to improve their current knowledge of the job and prepare for future job prospects. To this end, the subject matter of the course was also indicative of the reason a student might take a MOOC. For example, Kizilcec and Schneider [16] found that students in a humanities course might have taken the course out of curiosity, versus students in a social science or health-care-related course, who might have taken the course for career benefits [5].

Others have investigated whether students' reasons for enrolling in a MOOC impacted their behavior during the course and whether or not students completed the course. For example, de Barba et al. [7] found that students' motivation and their interests were related to how they engaged with the course's quizzes and videos. They also investigated how motivation—either intrinsic motivation or situational interest—was related to a student's final grade in an introductory economics MOOC. Others, however, observed no relation between student motivation and the grades earned in MOOCs [3]. On the other hand, Pursel and co-authors [23] found that students who had the intention to be an active participant in a MOOC had higher odds of completing the MOOC. In other words, those who stated they were motivated to finish the MOOC were actually more likely to do so.

We also note that a few studies have investigated students' reasons for enrolling in MOOCs by analyzing open-ended survey questions. For example, Robinson and colleagues [25] analyzed n-grams from the responses to a survey question that asked students how the course material was useful and how they planned to use the knowledge gained from the course. Using regularized regression, they found students whose answers included words that indicated a plan to readily apply the knowledge gained from the course, and expected to use the skills learned from the course in a vocational setting, were more likely to earn a certificate than students whose responses indicated an interest in obtaining formal recognition. In another investigation of open-ended survey responses, Crues et al. [6] found that students' reasons for enrolling in a MOOC clustered into four interpretable reasons, and some of the reasons were related to actively engaging in portions of the course; however, these reasons were not statistically related to remaining engaged in the course overall. In general, much more can be learned from students' motivations and goals for enrolling in MOOCs, and this new knowledge can be utilized to further an understanding of students who take these courses.

2.2 Role of Gender and Age in MOOCs

MOOCs can provide informal experiences for students, with few barriers and no requirements for enrollment, but this also leaves MOOCs without traditional educational data about student background characteristics. However, there have been several studies that have explored the relations among student characteristics, enrollment patterns, and behavior in MOOCs. In this paper, we focus on the relation between two background characteristics—gender and age—in understanding reasons for enrolling and behavior in MOOCs. We have chosen to examine gender because of MOOCs' great promise to offer educational experiences to all, which has particular importance for women, who often-

times have fewer educational opportunities than men. In addition, men and women might have different patterns of enrollment in different courses, and having this information could be vital for modifying and improving a course. We have also chosen to investigate age because older learners and younger learners might engage with MOOCs for very different reasons, and we want to document evidence on this issue.

With respect to gender, differences have been observed in whether males or females take a certain MOOC. Specifically, courses focused on science technology, engineering, and mathematics (STEM) tend to be dominated by male students [24, 10, 3]. For example, Breslow and co-authors [3] investigated "Circuits and Electronics" and found that 88% of the students who submitted an end-of-the-course survey were male. Women, more than men, are more numerous in other fields [20, 24]. And although men are more numerous in some STEM fields, medicine seems to be an exception: a course in medicine analyzed by Kizilced and Schneider [16] was overwhelmingly female—91% of students were female. We suspect that knowing the gender composition of the course is useful information to the instructor, especially if an instructor's goal is to attract more women or more men to the course.

The age of students in MOOCs has often revealed that students are young [5], with little variation between courses in different academic disciplines [16, 20]. Others, however, have found there to be a wide range of ages in classes (e.g., [3]), and that age varies based on geography [10]. The disparate findings on the age of MOOC students suggests that the relation between student age and participation in a MOOC is still murky and further research could be done to clarify this relationship.

Students' ages and genders have often been found to share (at best) a weak relationship with their reasons for enrolling in a MOOC. With respect to gender, Crues and colleagues [6] observed that students' reasons for enrolling in a computer science MOOC and gender did not share a significant statistical relationship.

Some have reported that females selected more reasons for enrolling in a MOOC on the Online Learning Enrollment Intentions (OLEI) scale than males [16]. In that study, reasons for enrolling in a MOOC were found not to be related to the age of a student. However, students who were using the MOOC to supplement their formal schooling were generally younger than students who did not indicate this reason for enrolling in the MOOC [16].

Although student gender and age have been found not to share a relationship with student reasons for enrolling in MOOCs, these background characteristics have been identified as sharing a relationship with student behaviors in MOOCs. For example, female students tend to spend more time viewing videos and completing assignments than males [24]. Although Swinnerton, Hotchkiss, and Morris [26] found that gender was not statistically related to the number of comments a student posted in a MOOC forum, others [24] found that females in non-science courses posted more inquiries in forums than males, but the opposite has been found to be true for science courses. Furthermore, it has been found that the reasons students gave for enrolling in a MOOC were related to their forum participation—men who enrolled to complement their career goals and women who did so to explore the content (e.g., they were curious about the course's subject matter) were more active in the forums than students who gave other reasons for taking the MOOC [6]. Findings have been inconclusive on whether gender shares a relationship with completing a MOOC: some investigations have found that gender shares a relationship with remaining persistent in a MOOC (e.g., [6]) or earning a certificate, depending on the course (e.g., [24]), while others have not observed this effect (e.g., [3, 20]).

Students' age has also been used to shed light on students' behavior in MOOCs. It was found that older students were more engaged with a MOOC than younger students; older students were found to have accessed digital course materials more frequently than younger students [10]; and older students were more active in the course forums than younger students [26, 10]. More generally, older students have been found to access more of the course materials than younger students [20, 10]. Similar to gender, there has been inconclusive evidence about whether age shares a relationship with success and completing a MOOC. For example, some have found that age was statistically related to grades (e.g., [10]) but others have not observed this effect (e.g., [3]. Still others have found that gender and completing a MOOC are not related [20].

In general, the literature has pointed to age and gender to be of interest in predicting enrollment and success in MOOCs, but the findings are not clear. Furthermore, we need to know more about why certain students enroll in some courses, and which of these reasons apply to MOOCs, in general, and which of these reasons only apply to particular MOOCS. Gaining insight on these issues is crucial for instructors and course designers to consider for attempting to improve courses. Thus, we conducted our investigation to provide more clarity on these issues.

3. METHOD

We used survey data to understand why students enrolled in one of five MOOCs offered on Coursera: 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). Students who enrolled in these courses were asked to submit a survey that asked about their background and expectations for the course, along with their age range and gender. The survey posed the questions, "Why are you taking this course? What do you hope to get out of it?" Students were able to enter an answer to both questions in one open-ended response. We call this the *reason* the student enrolled in the MOOC. We analyzed the responses to this survey to understand (1) why students enrolled in these MOOCs, (2) whether these reasons were related to specific courses or to the five MOOCs, in general, and (3) how reasons and courses were related to the students' background characteristics (gender and age).

Of the $N = 341523$ students enrolled in these MOOCs, $n = 37178$ responded to portions of the aforementioned sur-

vey; however, only $n = 12407$ students provided a reason that they enrolled in the course. As a result, these are the only students we will consider for analysis. In addition, because we used LDA to analyze the reasons that students enrolled, we removed non-English responses (using the `textcat` package in R [8]). This resulted in the total number of responses to be analyzed as $n = 11202$. The students who provided responses in English were spread throughout the five courses as shown in Table 1. The gender and age distribution for these courses is also displayed in Table 1.[2]

After we removed non-English responses, we prepared the text for analysis using the `tm` package in R [19]. Before we did any text pre-processing, there were 11058 unique words in the set of English responses. We removed stop words, punctuation, and numbers, while also transforming all characters to lower case and stemmed the terms using the Porter stemming algorithm [22]. Additionally, the term frequency-inverse document frequency (tf-idf) scores were computed for the collection of reasons. We removed terms that had tf-idf scores at or below the tenth percentile, because these terms might include more noise in the text data. After completing these pre-processing steps, we had 9952 unique terms in the set used to model these responses.

To model these responses, we used Latent Dirichlet Allocation (LDA) [2], which is a type of unsupervised topic model. Topic models are probabilistic models, which assume that a collection of documents follow an underlying latent distribution [2, 12]. LDA is a well-suited method for this problem because the reasons students gave do not have a label attached to them, and our goal was to explore the relations between reasons and MOOCs. Specifically, the LDA model is defined as

$$p(\theta, \mathbf{t}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{i=1}^{I} p(t_i|\theta) \cdot p(w_i|t_i, \beta), \qquad (1)$$

where θ is the topic mixture, t is the number of topics I an LDA model assumes, w is the collection of words used to fit the model, α is a vector of length t, and β is a matrix of word probabilities [2]. To estimate these models, various estimation strategies have been proposed. One approach is variational expectation maximization (VEM) [2]; however, the starting values of the algorithm are non-trivial which could result in finding local, versus global, maximums [9, 2, 13]. To combat this problem, Gibbs sampling has been proposed to estimate the unknown parameters for LDA, and identifies these parameters faster than other algorithms [9]. Before estimating an LDA model, however, one must specify the number of topics, t.

To determine the number of topics in the collection of reasons, we used the strategy proposed by Griffiths and Steyvers [9], which was implemented using the `ldatuning` package in R [21]. After estimating LDA models where the number of topics was $I = \{10, 11, 12, ..., 35\}$, we found 26 topics was close to the maximum of the metric proposed in [9]; thus we fit an LDA model with 26 topics. We show the metric's behavior versus a subset of the number of fitted topics in Figure 1.

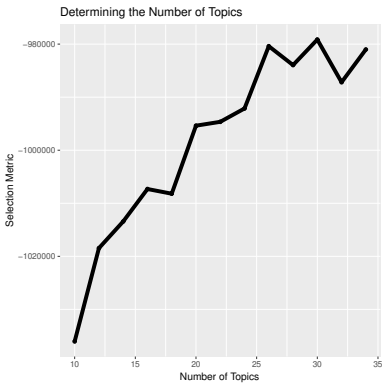


Figure 1: Plot of model fit metric in [9] versus the number of topics. When the number of topics is 26, the metric is near the maximum.

Once we determined $t = 26$, the LDA model was fit using the `topicmodels` package in R [14], where we used Gibbs sampling with 500 random starts and 5000 iterations, and the first 1000 iterations were discarded for burn-in. The final model selected was the one that had the highest posterior likelihood, and then we assigned each student to one of the 26 topics. This was done by computing the posterior probability from the LDA model, and each student's reason was assigned to whichever topic had the highest probability.

After we assigned each student to one of the topics, we cross-classified students by the topic to which they were assigned from the LDA model and the course in which they were enrolled. To test whether there was a statistical relationship between the topics and the particular course they were enrolled in, we used the χ^2 test of independence. We do not offer direct interpretations of the topics because it is difficult for humans to identify topics from a given set of terms from a topic model [4]. However, the potential relationship between courses and reasons lends itself to correspondence analysis, so we further analyzed this two-way table by correspondence analysis using `FactoMinR` [17].

Correspondence analysis was used to represent the association between reasons and courses using the data in Table 2. Plots of the estimated scores for the topics (rows) and courses (columns) represent the dependency in the table. The method can determine which reasons differentiate or are unique to particular courses and which reasons do not distinguish between the courses (i.e., the reason is common to all courses).

To investigate whether the relationships between topic and course was mediated by background characteristics, specifically gender and age, we fit log-linear models (using maximum likelihood estimation) to three-way contingency tables of topic by course by background characteristic. Our modeling strategy started with a complex model, and then we sought to find the most parsimonious model that yielded a good representation of the data. Specifically, we started by

[2]Students were able to identify as male, female, or neither of these. After filtering out students who did not provide a reason for enrolling or an answer in English, all remaining students identified as either male or female.

Table 1: Distribution of students enrolled in the five MOOCs by gender and age.

Course	Students		Gender		Age Group						
	Total	Complete Survey	Males	Females	≤ 17	18-24	25-29	30-39	40-49	50-59	≥ 60
Android	189334	4656	3589	1067	112	1055	980	1234	684	393	198
Ochem	38526	784	440	344	12	193	185	193	84	72	45
Subsistence	23854	729	312	417	3	103	161	196	97	91	78
Sustainability	76886	4199	1889	2310	17	520	917	1116	641	527	461
Elearning	12923	834	357	477	1	24	74	224	239	185	87

modeling the relationship using

$$\log \mu_{ijk} = \lambda + \lambda_i^b + \lambda_j^c + \lambda_k^t + \lambda_{ij}^{bc} + \lambda_{ik}^{bt} + \lambda_{jk}^{ct} + \lambda_{ijk}^{bct}, \quad (2)$$

where i corresponds to the levels of the background characteristics b (i.e., male and female for gender, or the 7 age groups), j corresponds to the courses, c, and k corresponds to the most probable topic, t, from the LDA model. Note then that μ_{ijk} is the number of students in cell ijk in the three-way contingency table. The analyses for gender and age were carried out separately. Once a model was chosen, we further studied the nature of the associations found in the data.

When using log-linear models, a Poisson distribution is typically assumed for the distribution of counts; however, we suspect that there was more heterogeneity within combinations of topic, course, and background than is predicted by a Poisson distribution (i.e., the data exhibit "over-dispersion"). To deal with this we used a negative binomial distribution in our log-linear models. Our conjecture that data were over-dispersed was confirmed. The dispersion parameter was large relative to its standard error and the negative binomial models yielded much better goodness-of-fit statistics. In all models and further analyses, we report the log-linear model and test statistics using a negative binomial distribution.

4. RESULTS

We first note that there were differences in student background characteristics across these five courses. From Table 1, we can see that there were more males in Android and Ochem, but more females in Subsistence, Sustainability, and Elearning. In general, there were few students aged 17 or younger in these courses. Most students were in the middle age groups. We used a likelihood ratio statistic of independence assuming a negative binomial distribution to test whether age and gender shared a statistical relationship, without respect to courses. The marginal relationship between gender and course was statistically significant ($X^2 = 10.03$, $df = 4$, $p = .03$), and the relationship between age and course was also statistically significant (i.e., $X^2 = 38.49$, $df = 24$, $p = .03$). Thus, we have evidence to believe that age and gender are statistically dependent with respect to who enrolls in these courses.

Table 2 defines the general topic model, where the five most probable words in each topic are listed with each topic and the number of student responses for each topic are displayed for each course. Note that the topics are ordered in an arbitrary manner.

To test whether there was a significant association between being enrolled in a specific course and assignment to a specific topic, we used a X^2 test of independence. Unsurprisingly, this test revealed a dependent relationship between topic and course (i.e., $X^2 = 12570$, $df = 100$, $p-$value $< .001$).

Furthermore, to gain insight into the nature of the relationship between topic and course, we performed a correspondence analysis. The first two dimensions account for 68.91% of the total inertia, which is a measure of the amount of association in the data (i.e., how much the data deviate from expectations under independence). The category scale values from the first two dimensions of the correspondence analysis are plotted in Figure 2. Greater distances between points for the courses indicates that there are greater differences in their profiles, with respect to the topics (a profile corresponds to the conditional distribution of topics, given course). Likewise, greater distances between points for the topics indicate greater differences in the profiles with respect to the courses.

The course points for Subsistence, Sustainability, and Elearning are close together, which indicates that these three courses have similar profiles with respect to the topics. These three courses are the least distinguishable in terms of the topics. The Android and Ochem points are far from each other and far from the other three courses, which indicates that these courses have considerably different profiles with respect to the topics and are quite distinct.

Although the absolute distances between the course and topics points are not meaningful, the relative distances between course and topic points are meaningful. For example, the points for topics (the reasons) 9, 19, and 20 are relatively close to Android, which means that these topics were given as a reason for taking Android more often than would be expected if topics and courses were independent. As can be seen in Figure 2, as we just noted, topics 9, 19, and 20 are relatively close to Android (most probable words: android, program, learn, develop, app), topic 2 is relatively close to Ochem (most probable words: chemistri, organ), topics 1, 10, 11, 15, and 17 are relatively close to Elearning (most probable words: understand, better, teach, onlin, world, way, work, current, interest, subject), topics 10, 17, and 25 are relatively close to Sustainability (most probable words: teach, onlin, interest, subject, studi, field), topics 7, 12, and 23, are relatively close to subsistence (most probable words: sustain, environment, market, social, can, chang), topics 7, 22, 23, 26 are all relatively close to Subsistence and Sustainability (most probable words: sustain, environment, sustain, system, can, chang, sustain, sustainability), and topics 10,

Table 2: Number of students matching each topic in the topic model, with distinctive words characterizing each topic.

Topic	Most Frequent 5 Words	Android	Ochem	Subsistence	Sustainability	Elearning
1	understand,better,hope,abl,gain	190	28	69	340	69
2	chemistri,organ,school,take,chemistry	76	466	6	117	20
3	take,course,the,also,reason	169	16	26	171	36
4	one,know,think,need,realli	164	24	18	234	28
5	use,make,can,like,idea	321	6	19	96	37
6	will,help,hope,give,think	182	25	24	178	30
7	sustain,environment,issu,sustainability,topic	40	2	23	406	14
8	knowledg,improv,skill,field,knowledge	255	25	43	277	49
9	android,program,app,apps,comput	1029	1	4	9	5
10	teach,onlin,educ,elearn,technolog	67	11	12	102	280
11	world,way,can,find,peopl	79	8	37	157	14
12	market,social,develop,work,countri	43	1	244	100	12
13	want,learn,know,just,curious	185	14	20	138	15
14	time,coursera,class,enjoy,great	84	52	11	170	18
15	work,current,project,area,compani	74	5	27	154	33
16	learn,new,someth,want,thing	210	8	17	111	32
17	interest,subject,area,view,point	78	6	35	219	25
18	like,interest,look,topic,see	123	5	19	112	17
19	learn,develop,want,development,basic	221	7	7	43	20
20	android,app,develop,creat,mobil	828	1	2	4	0
21	get,hope,job,good,field	85	8	7	90	14
22	sustain,system,food,product,energi	12	6	16	225	4
23	can,chang,sustain,futur,human	7	2	12	292	15
24	year,time,ive,now,tri	93	32	13	83	9
25	studi,field,degre,research,master	24	23	15	175	32
26	sustain,sustainability,concept,practic,need	17	2	3	196	6

11, 15, 17, and 25 are relatively close to sustainability and Elearning (most probable words: teach, onlin, world, way, work, current, interest, subject, studi, field). The topics in the center of the figure (i.e., 3, 4, 6, 8, 13, 14, 16, 21, and 24) are those that do not differentiate the courses and are given as reasons for all courses (probable terms include take, course, one, know, will, help, knowledg, improv, want, learn, time, coursera, learn, new, get, hope, year, time). Next, we consider how the student background characteristics are related to the reasons and the courses.

4.1 Gender, Reasons, and Courses

We fit log-linear models to understand the relationship between student gender, the topic a student was assigned to from the LDA model, and the course they took. The homogeneous association model (all 2-way interactions, but not the 3-way interaction from Equation 2) yielded an excellent representation of the data (i.e., the likelihood ratio goodness-of-fit statistic was $X^2 = 49.186$, $df = 100$, $p = .99$). Among the three possible conditional independence models (i.e., only two two-way interaction in equation 2) fit to the data, only the model where topic and gender are independent given course gave a good representation of the data (i.e., $X^2 = 16.318$, $df = 25$, $p = .91$).

Given that the topic and gender were conditionally independent given course, we could collapse over gender to study the topic by course relationship and collapse over topic to study the relationship between gender and courses [1]. We have already described the relationship between course and topic based on the correspondence analysis. Figure 2 described both males and females; in other words, there are no difference between males and females in terms of the dependency between courses and topics.

To study gender by course dependency, we refer to the middle of Table 1. We found, using a negative binomial distribution, that gender and course were dependent. Table 3 contains Haberman residuals from the independence model. We chose to use Haberman residuals, which are related to standardized Pearson residuals, because Haberman residuals are distributed $N(0, 1)$, whereas the distribution of Pearson residuals is $N(0, < 1)$ [1]. The Android course was the only course were there was a noticeable difference between males and females. The males enrolled in the Android course more than expected and females enrolled far less than expected.

4.2 Age, Reasons, and Courses

Similar to the analysis for gender, we fit log-linear models (again, using the negative binomial distribution) to the topic-by-course-by-age, 3-way table. As before, the homogeneous association model yielded an excellent representation of the data (goodness-of-fit likelihood ratio test statistic $X^2 = 470.355$, $df = 600$, $p = .99$). None of the conditional independence models yielded an acceptable goodness of fit to the data. We were not able to collapse over any of the variables to describe the association between pairs of variables [1], so we further examined the partial tables (i.e., the relationship between age and topic, age and course, and course and topic) to describe the association between pairs of variables with an emphasis on the topic-by-course interaction.

To further explore the relationship between age group, the topic from the LDA model, and the course a student took, we

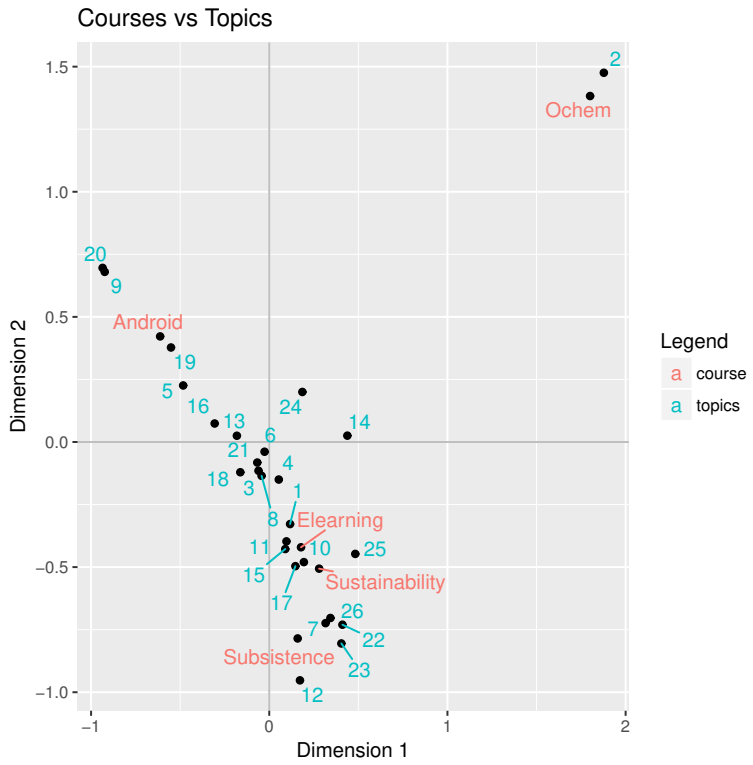


Figure 2: Topics and courses from Table 2 projected in the first two dimensions of correspondence analysis (association between topics and reasons for enrolling).

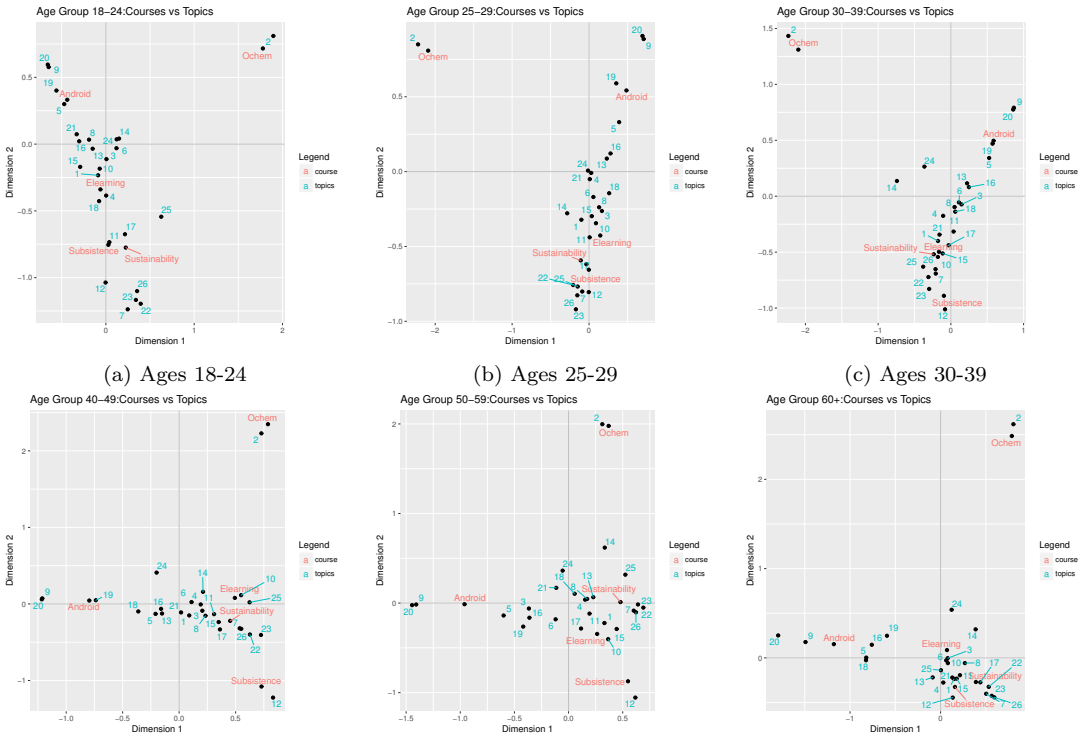
Table 3: Haberman residuals from independence using the Negative Binomial distribution.

	Course				
Gender	Android	Elearning	Ochem	Subsistence	Sustainability
Male	-2.8035	1.13055	-0.3422	1.12843	0.90563
Female	2.80349	-1.1305	0.3422	-1.1284	-0.9056

used correspondence analysis where we completed a separate analysis for each age group. We include six correspondence analysis plots for the first two dimensions in Figure 3 for all age groups except the youngest students, because there were very few students in this category ($n = 145$). Table 4 gives the proportion of total inertia accounted for by the first two dimensions of the plots in Figure 3.

Generally, Ochem is far from the other courses and, relative to the other courses, is far from all but one topic (topic 2, where the most probable words are chemistri and organ). Likewise, Android is relatively far from other courses as shown in Figure 3. In all of the plots, we observe that topics 9 and 20 are quite close to Android, which is intuitive given that the most probable words for these topics are android, program, app, and develop. Furthermore, across the different age groups, topic 19 is relatively close to Android, where the most probable words are learn and develop. Across all of the age groups in Figure 3, topic 11 is generally close to Subsistence, where the most probable words are market, social, and develop. We see that for most students, topic 10 is quite close to Elearning. The most probable words for this topic are teach, onlin, educ. In most of the plots in Figure 3, topic 17 is generally close to Sustainability, and the most probable words for this topic are interest, subject, and area. For the other topics, it is more difficult to establish a clear pattern across the different age groups. In other words, many of the topics do not consistently differentiate the courses from one another, and thus, are reasons given for all of the courses.

To further understand the relationship between students' age and the topic they were assigned to, given the course they took, we considered the Haberman residuals of the partial tables. That is, we considered the residuals for five 2-way tables, where each table corresponded to one of the MOOCs, and the rows and columns corresponded to the topics and age groups. Because Haberman residuals follow the standard normal distribution, any residual with an absolute value of two or greater is of note. Out of the 910 residuals, there were eight residuals with an absolute value greater than 2 for Android, 13 for Elearning, 8 for Ochem, 12 for Subsistence, and 4 for Sustainability. The large residuals in this case were generally for the two youngest age groups (i.e., students 24 years old and younger) or the two oldest age groups (i.e., students 50 and older). This suggests that, given the course a student took, we saw students in these four age groups were assigned to topics much more or much less than expected. This means some younger and older stu-



(a) Ages 18-24 (b) Ages 25-29 (c) Ages 30-39

(d) Ages 40-49 (e) Ages 50-59 (f) Ages ≥ 60

Figure 3: Correspondence analysis plots for all age groups except the youngest.

Table 4: Inertia accounted for by the first two dimensions, as shown in Figure 3.

Age Group	Inertia from first two dimensions
18-24	74.50%
25-29	75.89%
30-39	69.43%
40-49	63.55%
50-59	63.54%
≥ 60	65.60%

dents take courses for reasons that were not expected, when we account for the course they took.

Additionally, we considered the partial tables to understand the relationship between age group and the course a student took, given the topic they were assigned to from the LDA model. We examined the Haberman residuals for 26 tables—one for each topic; in this case, the rows and columns correspond to the age group and course a student took, and the cells contain the Haberman residuals. We found 45 out of 910 residuals with an absolute value of two or greater. Many of these larger residuals were for students in the two youngest and two oldest age groups. This suggests that students in these age groups took some of the courses much more or much less than expected, given the topic they were assigned to from the LDA model.

5. DISCUSSION

This investigation explored the reasons students gave for enrolling in one of five different MOOCs, and how these reasons related to the course students took, their gender, and their age. The five courses considered in this paper are from diverse academic disciplines and attract different groups of students.

Unlike some previous studies that have explored student goals for enrolling in MOOCs by asking them to select a reason from predetermined answer choices, students in this study specified their reasons for enrolling via an open-ended response. This afforded students the opportunity to provide more genuine responses, versus being forced to conform to a set of choices on a survey. As a result, we found 26 reasons students gave for enrolling in these MOOCs when using LDA. The number of topics for the LDA model, which must be specified, was derived empirically from the approach given in [9]. From this topic model, we observed that some students decided to enroll in a course for very specific reasons and we suspect that these specific reasons were related to the course content. This follows from the fact that some topics were very close to courses in the correspondence analysis; further support for this comes from the most probable words from each of these topics. On the other hand, some topics from the LDA model applied to all courses. These topics were those that were towards the center of the correspondence analysis plots. When examining the most probable terms for these topics, we found very general terms that did not have an apparent relationship to one of the five courses we considered.

We also examined whether students' gender or age were related to the courses they took and the reason they enrolled in the course. We first considered whether a students' gender,

the course they took, and the topic they were assigned to from the LDA model were statistically related. Our analyses revealed there was not a 3-way interaction between these factors; however, our findings led us to analyze the relationship closely between topics and courses, and courses and gender. It was observed that gender did not mediate the relationship between topics and courses, thus, our findings about how the topics and courses are related is not different for males versus females. This finding is consistent with previous studies, which have found that, generally, the reason a student enrolls in a MOOC and their gender are not related (cf. [6], [16]). On the other hand, we found that there was a relationship between the courses students took and their gender. Some of the courses, such as those in the sciences, had more males, and those not in the sciences had more females. This finding parallels the enrollment patterns observed by Morris and colleagues [20].

We conducted a similar set of analyses to uncover the relationship between students' age, their topic assignment from the LDA model, and the courses they took. As when analyzing gender, we did not find a 3-way interaction between these three factors. Instead, we found statistically significant relationships between all of the 2-way interactions between these factors. To study the relationship between course and topic, given age group, we used correspondence analysis. Here, we found that one course, Ochem, and a reason related to enrolling for Ochem, were far from the other courses, and the other four courses considered in this paper shared similar relationships with one another across age groups. To further understand the relationship between these three factors, we analyzed how age group and course, given their reason for taking the course, were related. When considering this relationship, we generally observed that students in the younger and older age groups enrolled in some of the courses more than expected. When more closely considering the topic from the LDA model and student age group, given the course a student took, we often found students in the younger and older age groups gave topics more or less than we would expect. This suggests that the students in this study who are in the two youngest and two oldest groups take courses and give reasons we might not expect.

Implications for course design: The finding that there is an age and gender dependence with respect to who enrolls in the courses may be interpreted as follows: Course designers could increase course effectiveness by including potentially age-relevant learning modules, such as a project or application focus for those in the degree-earning and job-seeking ages and information or lecture focus for those outside these ages. Furthermore, while the dependent relationship between reason and course suggests the obvious—learners are in different courses for different reasons—it could also be construed to mean that specific changes, such as the optional learning modules mentioned above, could improve course effectiveness.

In general, the approach in this paper can be used to characterize students' reasons for enrolling in MOOCs and subsequently to improve MOOCs. For example, students who feel isolated from their peers are often dissatisfied with their online courses. One of the potential ways of improving this situation could be to provide ways for learners who enroll to find community to connect with others who share this goal, thereby potentially ameliorating their isolation. In addition, instructional designers could help learners customize their learning experience if they knew how learners with different reasons for enrolling engaged differently with a course. For example, content choices can be categorized as being introductory and advanced, and multiple learning paths could be suggested at the outset, allowing more advanced students to jump to the appropriate content rather than have to wait or muddle through and be bored with the content that they have already mastered. As another example, those motivated to advance their job potential may be provided with assignments and projects that involve authentic work applications of the material, in contexts relevant to their particular situations. In general, understanding students' reasons for enrolling in a MOOC provides key information for improving the course and improving students' experiences with that course.

Future directions: Understanding reasons for MOOC enrollment is only one part of improving course effectiveness. Future studies in this direction should analyze how learners with different goals engage with a course in combination with their patterns of engagement while in the course, and how long they stay in the course, all towards improving learning experience for those participating in MOOCs.

6. CONCLUSION

We found that students take MOOCs for many different reasons. Although multiple-choice survey responses are useful to understand the reasons that a student might enroll in a MOOC, we found it is also feasible to use students' open-ended responses to questions that asked about why they were taking the course and what they hoped to learn. We found that some of the reasons students enrolled in these MOOCs were course specific, while others showed a general interest in learning or taking a MOOC. By examining *why* students take MOOCs, we can develop a greater understanding of what students might want when they take a MOOC. If the reasons a student takes a MOOC are more thoroughly understood, it could help explain why MOOCs have such high attrition rates and provide insight to ameliorate this issue, ultimately improving retention and learning.

7. REFERENCES

[1] A. Agresti. *Categorical data analysis.* Wiley-Interscience, 3rd. edition, 2013.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[3] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton. Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, 8, 2013.

[4] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pages 288–296, 2009.

[5] G. Christensen, A. Steinmetz, B. Alcorn, A. Bennett, D. Woods, and E. J. Emanuel. The MOOC phenomenon: Who takes massive open online courses and why? 2013.

[6] R. W. Crues, G. M. Henricks, M. Perry, S. Bhat, C. J. Anderson, N. Shaik, and L. Angrave. How do gender, learning goals, and forum participation predict persistence in a computer science MOOC? *ACM Transactions on Computing Education*, 2018.

[7] P. G. de Barba, G. E. Kennedy, and M. D. Ainley. The role of students' motivation and participation in predicting performance in a MOOC. *Journal of Computer Assisted Learning*, 32(3):218–231, 2016.

[8] I. Feinerer, C. Buchta, W. Geiger, J. Rauch, P. Mair, and K. Hornik. The textcat package for n-gram based text categorization in R. *Journal of Statistical Software*, 52(6):1–17, 2013.

[9] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[10] P. J. Guo and K. Reinecke. Demographic differences in how students navigate through MOOCs. In *Proceedings of the first ACM conference on Learning@Scale*, pages 21–30. ACM, 2014.

[11] K. F. Hew and W. S. Cheung. Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. *Educational Research Review*, 12:45–58, 2014.

[12] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[13] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001.

[14] K. Hornik and B. Grün. topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30, 2011.

[15] J. P. Howarth, S. D'Alessandro, L. Johnson, and L. White. Learner motivation for MOOC registration and the role of MOOCs as a university 'taster'. *International Journal of Lifelong Education*, 35(1):74–85, 2016.

[16] R. F. Kizilcec and E. Schneider. Motivation as a lens to understand online learners: Toward data-driven design with the OLEI scale. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 22(2):6, 2015.

[17] S. Lê, J. Josse, F. Husson, et al. FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 2008.

[18] M. Liu, J. Kang, and E. McKelroy. Examining learners' perspective of taking a MOOC: Reasons, excitement, and perception of usefulness. *Educational Media International*, 52(2):129–146, 2015.

[19] D. Meyer, K. Hornik, and I. Feinerer. Text mining infrastructure in R. *Journal of Statistical Software*, 25(5):1–54, 2008.

[20] N. P. Morris, S. Hotchkiss, and B. Swinnerton. Can demographic information predict MOOC learner outcomes. *Proceedings of the EMOOC Stakeholder Summit*, pages 199–207, 2015.

[21] M. Nikita. *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters*, 2016.

[22] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[23] B. K. Pursel, L. Zhang, K. W. Jablokow, G. W. Choi, and D. Velegol. Understanding MOOC students: Motivations and behaviours indicative of MOOC completion. *Journal of Computer Assisted Learning*, 32(3):202–217, 2016.

[24] J. Qiu, J. Tang, T. X. Liu, J. Gong, C. Zhang, Q. Zhang, and Y. Xue. Modeling and predicting learning behavior in MOOCs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 93–102. ACM, 2016.

[25] C. Robinson, M. Yeomans, J. Reich, C. Hulleman, and H. Gehlbach. Forecasting student achievement in MOOCs with natural language processing. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 383–387. ACM, 2016.

[26] B. Swinnerton, S. Hotchkiss, and N. P. Morris. Comments in MOOCs: Who is doing the talking and does it help? *Journal of Computer Assisted Learning*, 33(1):51–64, 2017.

[27] S. Zheng, M. B. Rosson, P. C. Shih, and J. M. Carroll. Understanding student motivation, behaviors and perceptions in MOOCs. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 1882–1895. ACM, 2015.