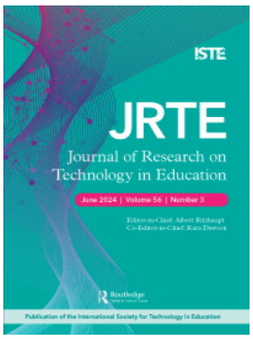

Routledge
Taylor & Francis Group

Journal of Research on Technology in Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ujrt20

Analyzing learners engagement in a micromasters program compared to non-degree MOOC

Farahnaz Soleimani, Jeonghyun Lee & Meryem Yilmaz Soylu

To cite this article: Farahnaz Soleimani, Jeonghyun Lee & Meryem Yilmaz Soylu (2024) Analyzing learners engagement in a micromasters program compared to non-degree MOOC, Journal of Research on Technology in Education, 56:3, 233-247, DOI: 10.1080/15391523.2022.2121342

To link to this article: https://doi.org/10.1080/15391523.2022.2121342

JOURNAL OF RESEARCH ON TECHNOLOGY IN EDUCATION
2024, VOL. 56, NO. 3, 233–247
https://doi.org/10.1080/15391523.2022.2121342

Analyzing learners engagement in a micromasters program compared to non-degree MOOC

Farahnaz Soleimani, Jeonghyun Lee and Meryem Yilmaz Soylu

Center for 21st Century Universities, Georgia Institute of Technology, Atlanta, Georgia, USA

ABSTRACT

This study aimed to understand the relationship between course activities and learning progress among students enrolled in the MicroMasters certificate program offered in an affordable MOOC-based learning platform. In order to capture the relationship, the differences between the engagement patterns of learners in the MicroMasters program compared to a non-degree MOOC were examined by utilizing machine-learning (ML) techniques in the clickstream database. The ML analyses revealed discrepancies in activity patterns and progress rates of students enrolled in MicroMaster and MOOC courses. The findings can further support optimizing the program's design to enhance learners' engagement and improve the overall completion rates.

ARTICLE HISTORY
Received 6 March 2022
Revised 6 June 2022
Accepted 29 August 2022

KEYWORDS
MicroMasters; online higher education; Massive Open Online Course (MOOC); clickstream data; learning analytics; student engagement

Introduction

MOOC-based micromasters credential program

Over the past few years, MOOC providers have started offering various types of online degrees and professional certificate programs. MicroMasters programs refer to a series of high-quality MOOC-based graduate-level courses provided by universities through the edX platform which propose learning of a certain topic with application in industry. The Massachusetts Institute of Technology (MIT) introduced the pioneer of these programs in 2014 as Supply Chain Management; afterwards, edX has delivered more than 50 MicroMasters programs to help learners promote their profession or pursue an advanced degree. In addition, MicroMasters programs have attracted interest and high enrollment rates, not only because they are flexible and affordable, but also completion of these programs increases the chances of getting admission to graduate programs, obtaining job promotions and related job offers (Hollands & Kazi, 2019; Ingolfsdottir, 2016). For example, Littenberg-Tobias and Reich (2020) found that lower costs, as well as a flexible and shorter timeframes of MIT's supply chain management, were important factors that encouraged students to apply for this professional credential program. However, the authors raised a concern that this type of program may not necessarily improve educational equity in access for under-represented demographic groups, but rather benefit already high-achieving groups such as male and advanced degree holders.

Despite the cultivation of a deep understanding of academic engagement and learning metrics among learners situated in traditional MOOCs over the last decade, research on the recently emerged MicroMasters program learners' behaviors and learning outcomes is relatively new. The most reported benefits among those who complete the programs were learning something new

CONTACT Farahnaz Soleimani soleimani@gatech.edu Center for 21st Century Universities, Georgia Institute of Technology, Atlanta, GA, USA
Supplemental data for this article can be accessed online at https://doi.org/10.1080/15391523.2022.2121342.

and improving current job performance. In addition, 7% of the total number of completers reported that the MicroMasters program certificate had improved their application to the related Master's degree (Dalipi et al., 2018). Revealed by Ponce-Cueto and Caplice (2019) and Cabral et al. (2017), MicroMasters programs are more than a set of individual MOOCs that have distinguished, unique characteristics compared to both the online and in-person traditional programs, thereby, more studies are needed to explore the particular properties of similar programs in light of designing and delivering the course contents carefully since the same common practices cannot always be applied to MicroMasters.

However, there exist only a few studies that directly compared learners' experiences between professional and other types of MOOCs. Yet, it is noteworthy that Gregori et al. (2018) examined three types of MOOCs (i.e., conventional, formal, and professional) to understand how online learner support is implemented. According to the results from their semi-supervised learning model, maintaining learners' engagement during the second quartile of the course as well as enhancing teacher presence played an important role across different MOOC types in predicting learners' completion behaviors. In this study, we seek to investigate the extent to which learners' patterns of navigating or interacting with the MOOC courseware and content would differ between those learners from the professional MicroMasters certificate program and those from conventional MOOCs that are accessible to anyone at any time.

Application of machine learning on MOOC data

Machine learning (ML) approaches have been extensively applied for the dropout prediction of the users (Moreno, 2019). In addition to dropouts, researchers utilized ML approaches to study long-term knowledge retention analyses (Davis et al., 2018), variations in the learning process based on participatory behavior and temporal learning paths (Rizvi et al., 2020), and evaluation of fairness and bias among support recipients (Lee & Kizilcec, 2020). Clickstream data containing the frequency of users' interaction, links opened, and videos watched were analyzed in similar studies to understand various phenomena in MOOCs. Nevertheless, researchers have not followed a standard in selecting the variables since there is no unification of clickstream data. This variation was noted as a binary classification problem.

In this regard, Al-Shabandar et al. (2017) analyzed both clickstream and demographic information to predict learning outcomes in MOOCs. The clickstream dataset included accumulated counts for video lecture views, access to assignments, and posts in discussion forums, while the demographic data contained students' age, reported gender, and educational background. With running different ML approaches, researchers uncovered a strong relationship between earning a certificate and features of clickstream data. Moreover, Sweeney et al. (2016) aimed to predict student success in each class they enrolled in every semester based on data including students' transcripts and information regarding instructors and courses. One of the striking results of this study was the use of hybrid Factorization Machines (FM), Random Forests (RF) method to predict the success or failure of both new and returning students enrolling in both new and current courses.

Along similar lines, Makhlouf and Mine (2020) used data from a longitudinal study in which data was aggregated within the ASSISTments software. ASSISTments software has been used to improve middle schoolers' math skills by providing immediate feedback to learners' responses while evaluating their knowledge and reporting these results to the teachers. The dataset consisted of clickstream log files and other related variables, including but not limited to students' college admission and first employment after college. With a combination of two metrics, the Area Under the Receiver Operating Characteristic Curve (AUC) and the Root Mean Squared Error (RMSE), the researchers conducted two types of analyses called "problem-based approach" and "skill-based approach" to examine students' performance. Results of the study indicated that the skill-based model was a better predictor than the problem-based model in describing the failing students and the thriving ones.

Likewise, Kloft et al. (2014) created a prediction model with a psychology MOOC and reported that the number of days active and the number of sessions were good predictors of dropout rates. Correspondingly, Liang et al. (2016) utilized four different ML approaches (namely support vector machine, Logistics Regression, Random Forest, and Gradient Boosting Decision Tree) to predict dropout from the XuetangX platform with the 40 days of student learning activities data. They used various user, course, and enrollment features to build a dropout prediction model and reported that Gradient Boosting Decision Tree was the best-performing approach.

These studies showed the critical role of utilizing ML approaches to investigate dropout and foreshadowing predicting factors such as educational background, posts in discussion forums, and accessed assignments. However, more studies are needed to examine students' progress, engagement, and performance in MOOCs. Besides, there is a lack of knowledge within the area of the MicroMasters programs, in particular investigating the clickstream data analysis in predicting learners' rate of progress throughout the programs.

Research scope of present study

Education researchers conceptualize student engagement as a multidimensional construct comprising behavioral, cognitive, and emotional (Deng et al., 2020). Learners' observable actions (e.g., note-taking, video activity), participation, and contribution to instructional activities (e.g., submitting an assignment) are associated with behavioral engagement (Deng et al., 2020). Since the actions are readily noticeable (via users' particular activities in MOOC platforms) and recordable, studies on learner engagement in MOOCs generally involved the behavioral dimension (e.g., Stöhr et al., 2019).

Given the emerging interest in these programs and to address the research gaps, this study investigates statistical properties of the registered learners in a MicroMasters program to better understand the characteristics of this population. Furthermore, this research evaluates learners' behavioral engagement (Fredricks et al., 2004) with the course contents, indicated by detailed records of their click activities in the online platform (i.e., clickstream data).

In this work, we apply various machine learning algorithms to predict learners' progress based on their online engagement in a MicroMasters program. The best model is selected as the one with the highest comparative prediction accuracy and the lowest cross-validation error. Moreover, the most significant clickstream activities and their level of importance in the prediction are analyzed. A similar procedure is followed to assess a regular MOOC course offered by the same institute. Comparison of the results helps to investigate the similarities and dissimilarities between the online engagements in courses that are offered in the context of a MicroMasters program and those that are known as regular individual MOOCs.

Finding the relation between the patterns of learners' activities and the program design forms the broader impacts of this study which is to investigate whether the course designs in such programs need to be modified to address further learner needs, or whether the designs need to be optimized to improve learner engagements. Foreshadowing the interactions between students and the online learning environment, this study can guide optimizing academic support based on a student-centered approach (De la Roca et al., 2018). In this regard, it can also provide insights into whether the rules and recommendations that apply to regular MOOC courses apply to the MicroMasters courses in terms of learners' online activities.

Materials and methods

MicroMasters program and regular MOOC settings

Our study is focused on investigating the MicroMasters Program in Analytics: Essential Tools and Methods, which offers three core graduate-level courses. These three courses came from the full Online Master of Science in Analytics (OMSA) degree program at a technology-focused

public research university in the U.S. The program was designed to help learners gain foundational knowledge of analytics as well as skills in applying essential analytics tools and techniques to various real-world contexts. It is a one-year program in which learners can access course content for free, while those who pay 500 USD are eligible to earn a verified certificate for each course. Students could earn the final program certificate if they earned a verified certificate for all three courses. The list of courses consists of Introduction to Analytics Modeling (ISYE6501x), Computing for Data Analysis (CSE6040x), and Data Analytics for Business (MGT6203x). To simplify presenting data and interpreting findings, Table A.1 in Appendix A lists assigned notations to these courses, offered in different semesters that are used hereinafter.

CSE6040x, a 15-week synchronous computer science course, provides learners with an introduction to essential programming principles and practices that are related to contemporary data analysis, data mining, and machine learning. The goal of this course is to develop proficiency in using modern analytical tools and techniques and applying them to real-world datasets for various mathematical or computational tasks. The course topics are grouped into three general units with each unit covering 5-6 component topics.

ISYE6501x is a synchronous 16-week course from the data analytics and statistics subject area. Learning objectives for this course are to understand fundamental analytics models as well as how and when to apply specific analytics models using analytical tools such as R, Python, and SQL. The lectures are divided into weekly sections with embedded modules that typically include 2-3 videos.

MGT6203x is a synchronous 10-week business and management subject course, which was designed to help students understand the scientific process of business data analytics and learn how to apply appropriate analytics methods to make informed business decisions. In contrast to ISYE6501x and CSE6040x which are conducted by a single instructor, MGT6203x is offered by the main instructor as well as two or three additional instructors who give lectures for given topic areas. The lecture videos are organized sequentially, which requires students to progress in a linear fashion.

The Introduction to Computing in Python course or CS 1301x is also provided by edX, as a regular MOOC, and it is designed to help learners to develop proficiency and back knowledge in computing and programming languages. This course is accessible to anyone regardless of having a computer science background or not. The course consists of seven modules in which instruction is delivered via a series of short videos.

Learner distribution and demographic trends

This section reports on student demographic information based on the results of an optional survey given during the course registration period. Overall, the average age of learners who participated in any of the three MicroMasters courses between 2017 and 2019, was 32.7 years old. Among 65,330 students who reported their country (88% of total), the majority of them were international students whose countries were other than the USA (72%; n = 47,252) whereas only 28% were USA students (n = 18,078). For gender distribution, 50,698 out of the total 74,010 students (69%) responded to the survey and, of those respondents, 26% reported that they were females (n = 13,072), 74% males (n = 37,349), and 1% other (n = 277). Additional information about students' levels of course engagement is provided in Appendix A (Table 2.A).

Overall, the three courses showed similar demographic trends over time (Figure 1). However, the percentages of female students from the Data Analytics for Business (MGT6203x) course tended to be slightly higher, compared to those from the other two courses: Introduction to Analytics Modeling (ISYE6501x) and Computing for Data Analysis (CSE6040x).

Regarding the level of education, students were generally highly educated. Most of the total 49,653 respondents (82%) reported that they were holding at least a Bachelor's degree (Table 1). Interestingly, more than one-third of the total respondents (34%) had already earned a Master's or professional degree, and even 4% indicated that they had completed a Doctorate level of

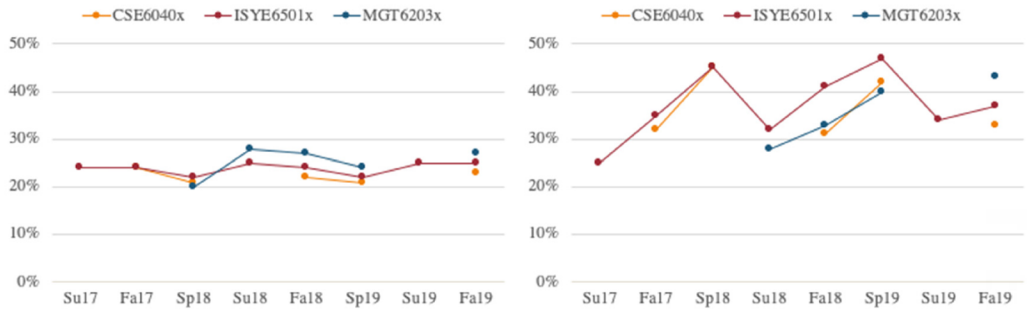

Figure 1. Percentage of (left) female students across time, (right) USA students across time.

Table 1. Distribution of Learners' Level of Education

	MicroMasters courses		
Level of Education	*CSE6401x*	*ISYE6501x*	*MGT6203x*
Doctorate	4.0%	4.6%	3.7%
Master's or professional degree	31.9%	36.8%	35.4%
Bachelor's degree	44.8%	45.1%	46.9%
Associate degree	2.9%	2.6%	2.6%
Secondary/high school	14.2%	9.1%	9.8%
Junior secondary/junior high/middle school	0.9%	0.5%	0.5%
Elementary/primary school	0.1%	0.1%	0.1%
Other Education	1.1%	1.0%	0.9%
No Formal Education	0.2%	0.2%	0.1%
Total	100.0% (n = 25,082)	100.0% (n = 36,063)	100.0% (n = 17,581)

education. Overall, the observed demographic trends with well-educated and male students as being dominant were similar to findings from another recent study which examined students who have completed at least one of five edX MicroMasters or five Coursera Specializations programs between January 2018 and October 2019 (Hollands & Kazi, 2019).

A total of 45,022 learners were enrolled in CS1301 during the year 2017. Of those learners who responded to the demographic survey at the beginning of the course (n = 37,257), 79% of the total were males, 20% females, and 1% others. The average age of the enrolled learners was 30.2 years. In terms of the reported level of education (total n = 36,283), a Bachelor's degree was the most common with 37% of the total; high school diploma next (25%); and Master's or professional degree (23%); 5% of the respondents Associate degree (5%); Doctorate degree (4%); middle school (3%); elementary school or no/other formal education (3%).

Clickstream database

In this study, we investigate the most common types of available clickstream data including video interaction, course navigation, and problem interaction events. Guided by previous research findings that suggest the advantage of using clickstream data to objectively measure student learning behaviors and therefore infer psychological processes that are closely associated with learning such as self-regulated learning (Li et al., 2020), the three specific clickstream event categories served as proxies for students' behavioral engagement with the online course materials. A description of these events is provided in Table 2. For example, the clickstream events "seq_next" and "seq_prev" refer to the attempts when a user selects the "next" and "previous" options in the navigation bar while "seq_goto" indicates when the user navigates to another unit.

The clickstream data is disaggregated by courses, and the contribution of events to the total students' engagements is compared for each course via a stacked bar chart (Figure 2). For an unbiased comparison, the students' engagements are compared for the courses offered in the

Table 2. Description of the Considered Clickstream Events

Clickstream data	Description*
load_video	Captures the event when the user fully renders the video to have it ready to play.
play_video	Captures the event when the user clicks on the video player's tab.
seq_next	Tracks the event when the user navigates from one unit to the next unit within the current subsection.
problem_check	Captures the event when the user successfully checks a problem, quiz or exam.
seq_prev	Tracks the event when the user navigates from one unit back to the previous unit within the current subsection.
seq_goto	Tracks the event when the user navigates to another unit within the subsection.
pause_video	Captures the event when the user when a user clicks on the video player's pause.
problem_save	Captures the event when the user saves a problem.
seek_video	Captures the event when the user goes to a different point in the video file.
link_clicked	Tracks the event when the user clicks on any hypertext link from the course content.
closed_captions_show	Captures the event when the user displays the closed captions.
show_transcript	Captures the event when the user displays the video transcript.
resume_course	Tracks the event when the user returns to the unit she/he was working on recently.

*https://edx.readthedocs.io/projects/devdata/en/latest/internal_data_formats/event_list.html.

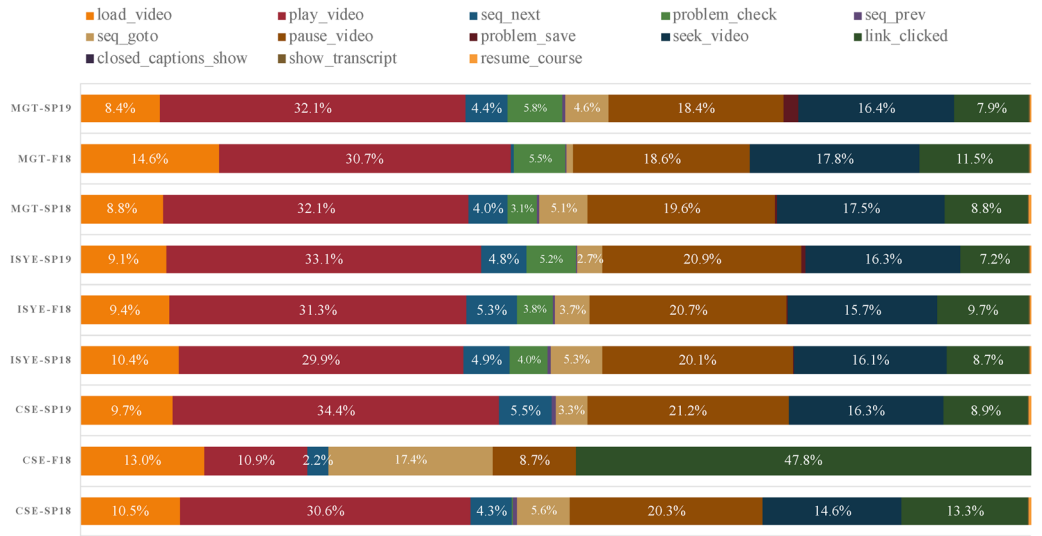

Figure 2. Distribution of activities in each course.

same semesters. As in the aggregated case, it is noted that the majority of events are related to interactions with video-type content. More specifically, it is observed that in the considered three courses almost 30% of clickstream activities relate to playing lecture videos while nearly 20% and 16.5% of the activities correspond to the actions of pausing and seeking videos. The only exception is CSE-F18, where the students clicked on the available links more often than they interacted with the video materials. This could be because the course was designed differently to give students access to the lecture videos and contents via additional links. Appendix B provides further details of our comparative analysis of the given clickstream data set.

Prediction of learner progress

To examine which one of the collected clickstream data corresponding to the learners' engagements with online content is indicative of their performance, prediction models are developed in this study. To this end, ML methods including multivariate linear regression, least absolute shrinkage and selection operator (LASSO), and tree-based methods are implemented. For each course, learners' cumulative individual activities during a semester are considered as the predictive variables to estimate learners' progress rate. Through developing predictions, the most influential activities are detected via feature selection techniques.

In the performance comparison (Tables C.1 and C.2 in Appendix C) of the applied ML approaches, the multivariate linear regression model had the lowest prediction power (e.g., 69.56%) among the generated ML models. It is observed that in most cases, other linear models, developed using either Forward stepwise or LASSO, had a slightly higher prediction power (e.g., 0.6% <1% for CSE6040x). More specifically, for the MicroMasters course ISYE6501x, the Forward Selection provides around 74% prediction accuracy, and the LASSO technique has a prediction accuracy of around 75%. Although Forward Selection and LASSO produced models with comparable statistical performance (i.e., prediction accuracy and cross-validation error) to the Full Linear Regression Model, they lower model dimensionality noticeably by using less than half of the total variables in their developed models. As a result, predictive models that are developed using such techniques are simple to implement in practice and easier to interpret.

Furthermore, the accuracy and validation errors are found to be improved by implementing tree-based algorithms including Bagging, Boosting, and Random Forest. This implies that the dataset involves inherent nonlinearity which the linear models are not able to capture. Particularly, in the case of the MGT6203x course, tree-based prediction models had about 13% higher prediction accuracy compared to the other methods. Also, for the other two courses, CSE6040x and ISYE6501x, tree-based models showed approximately 8% higher prediction accuracy. In addition, the tree-based algorithms were found efficient in reducing the cross-validation error by 55% on average using Bagging and Boosting algorithms and by 58% on average using Random Forest. For example, this reduction is around 50% (e.g., from 0.0140 to 0.007) for the CSE6040x course and around 65% (e.g., from 0.0076 to 0.0027) for the ISYE6501x course.

Overall, it was noted that Random Forest performed better than the other applied techniques. Moreover, Random Forest provides additional advantages such as overcoming the overfitting issue and ranking the level of importance of variables in the prediction. Figure 3a–c illustrates the contribution of clickstream activities in estimating students' progress rate. It is observed that using provided links, loading lecture videos, problem interaction, and showing captions are the most significant predictors in the completion of MicroMasters courses. As an example, the six most influential predictors for the ISYE6501x course are "link_clicked", "load_video", "problem_check", "seq_next", "closed_captions_show", and" show_transcript".

In order to distinguish the differences between the clickstream engagements of learners in the considered MicroMasters program and non-degree MOOCs, the progress rate of learners enrolled in a regular MOOC course, CS1301, was also predicted following a similar strategy. Similar findings were noted in terms of the performance of the applied algorithms. Tree-based regression models, particularly Random Forest, provided significant improvement to the predictive models (such as increasing prediction accuracy), while Forward Stepwise and LASSO have not advanced the linear full model performance. Notably, in both regular MOOC and MicroMasters courses, it was found that six input variables, detected as the most influential predictors, were sufficient to obtain optimum accuracy. Figure 3d presents the level of importance of the associated variables. In the following section, a comparison of the results is provided to elaborate on how differently the students' engagements affect their progress.

MicroMasters program completion analysis

An overview of the entire program is provided in this section. Around 4% of students took all three offered courses and among those who were enrolled in three courses, 25% successfully completed the MicroMasters program. A new predictive model is created to predict the average rate of progress throughout the program based on the average learner's activity in the three courses. Prediction accuracy of 74.5% is obtained by implementing the Random Forest algorithm, which found that the most important event types are "load_video" and "link_clicked". This is in agreement with the general findings from the analyses of individual courses.

Considering that students with an average progress rate above the average are more likely to complete the MicroMasters program, the online activities are compared for two groups of

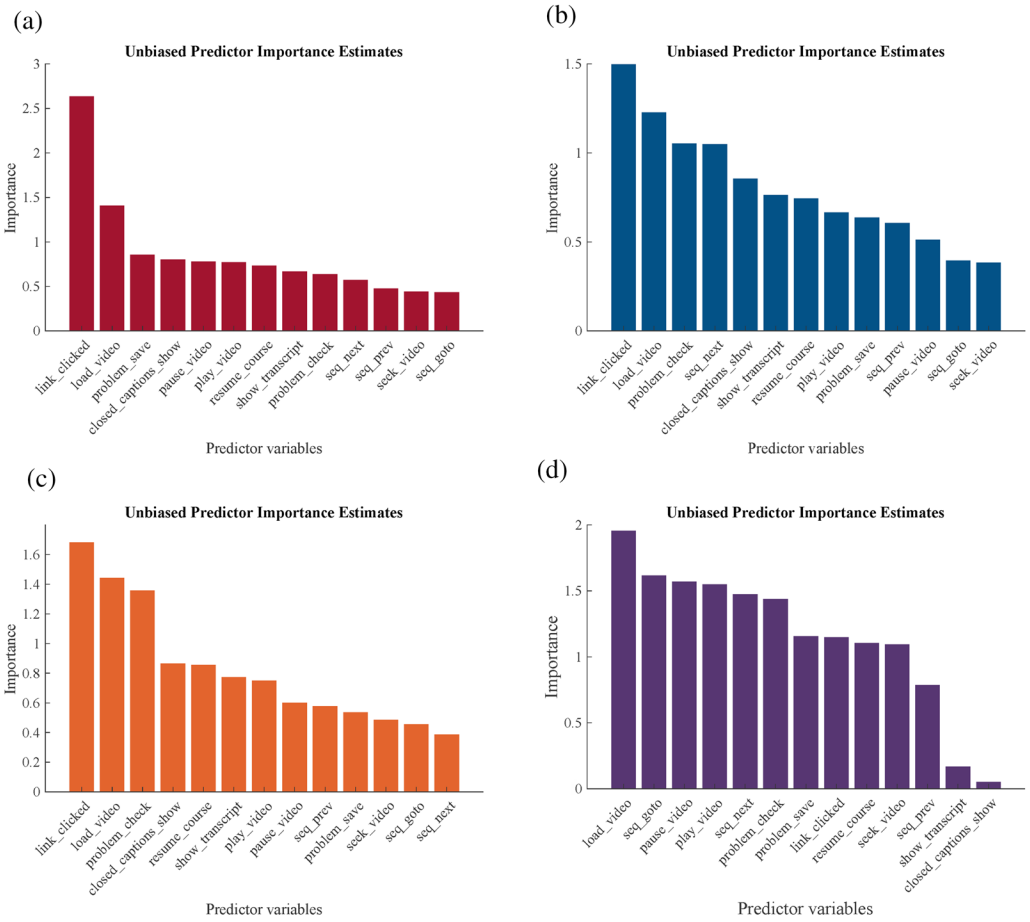

Figure 3. Significant clickstream activities identified by ML techniques for: a) CSE6040, b) ISYE6501x, c) MGT6203x, and d) CS1301.

students: I) G1: with an average progress rate equal to or higher than the average; II) G2: with an average progress rate lower than the average. As demonstrated in Figure 4, students can be differentiated based on their average activity pattern in three courses and their overall progress rate. Particularly for most events that were found to have a significant influence on the predictions (e.g., average load-video), the progress rate enhances as those activities increase.

Despite this observed positive trend, a particular pattern is not observed for the other event types with less impact on the learners' progress such as the average closed-captions-show. In addition, these two groups of learners (G1 and G2) are classified by applying classification algorithms and according to their influential clickstream data identified in the previous section. Using this approach, we aim to predict students with a progress rate higher and lower than the average. For this purpose, we classified students into two groups with the target labels (G1 and G2). In order to create the classification model and check the model performance, we randomly split our data into train and test subsets. Logistic Regression and Random Forest classifiers provide a classification accuracy of 84% and 85%, respectively. The cumulative average activities are compared in Figure 5 for the two groups of students to explore the differences in their source of activities. It is observed that students with a percentage of progress below 50% are on average less active in video-related events such as "play_video" and "seek_video" while they used other tools such as "seq_prev" and "resume_course" more frequently compared to the other more successful learners. This indicates this group paused the course more often.

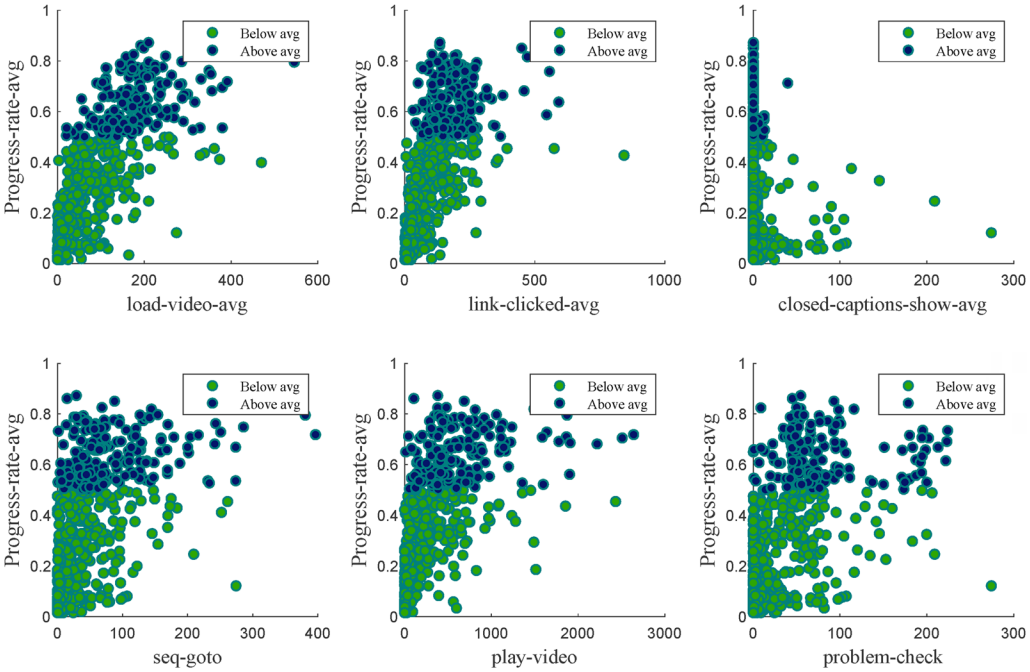

Figure 4. Average progress versus the average activities throughout the entire program.

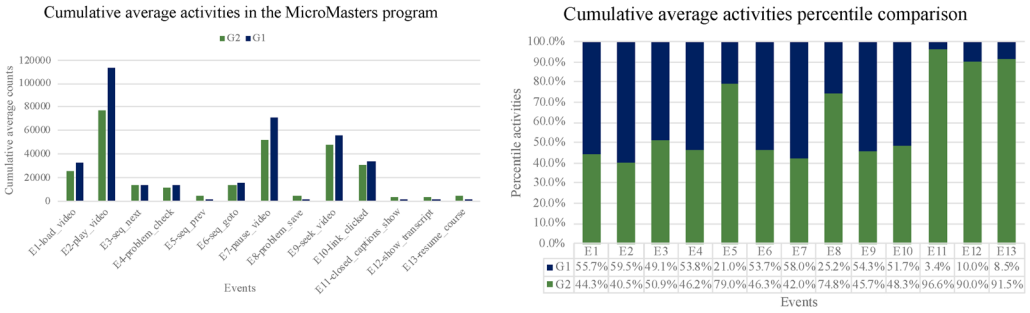

Figure 5. Engagement comparison of the classified learners.

Results

In this study, we applied various ML-based predictive modeling techniques to test the extent to which clickstream event data could predict how much progress students would make in the MOOC-based MicroMasters courses. By comparing the results from the feature importance analysis to Figure 2, engagement in both videos and problem sets, and not the individual total of either, predicts students' engagement. It is apparent that not the frequency, but the pattern of the learners' engagement or activity is a determinant of their level of importance in the predictions.

Regarding the best performed ML-based models, in all investigated MicroMasters and regular MOOC courses, the tree-based regression models accomplished significantly better than the linear regression models. Among the nonlinear approaches, due to slightly better overall performance and additional advantages such as measuring predictors' contribution, Random Forest evolves as a promising algorithm. The highest prediction accuracy (84%) was acquired for the ISYE6501x. The accuracy for CS1301, CSE6040x and MGT6203x was 80%, 77% and 70%, respectively.

The cross-validation errors of the regular MOOC across both linear and nonlinear ML-based models are lower than those of the MicroMasters courses. This indicates that the clickstream

activities of learners are more consistent in the regular MOOC since less variation was observed in this dataset. As a result, the predictions are more precise for the non-degree MOOC than the MicroMasters courses. The more consistent pattern of online engagement in the non-degree MOOC could be because of the fact that the learners are more familiar with the design of such courses since these courses have been offered by different institutions for a while, and they almost follow standard instructional designs and formats. However, learners keep exploring the options and new contents of micro-credential degrees which could be designed in different formats. This leads to more variations in the learners' engagements with the online course tools.

Although micro-credential programs such as the considered MicroMasters program in this study provide a new pathway for workers to advance their skills. They particularly face challenges in involving or expanding their access to the underrepresented groups, which is consistent with findings from Gregori et al. (2018). As the demographic comparison showed, the majority of enrollees in the MicroMasters program are mostly educated mid-career male professionals with an average age of 32 years. While we observed a slightly more diverse distribution of level of education and a lower average age from those learners who were enrolled in CS 1301, the issue of gender inequality of access to learning still appears to remain in the regular and low-stakes STEM MOOC. These findings warrant further investigation of the demographic distribution and its potential relationship with patterns of learning among learners from non-STEM and other types of MOOCs.

The predictive models included five to seven similar predictors for all analyzed courses. In this regard, the most significant clickstream events impacting learners' progress were:

- Accessing the provided links was the most critical event in the three MicroMasters courses; however, it was rated eighth for the regular MOOC. This implies that learners checked out most of the necessary course materials outside the MicroMasters platform. Nevertheless, accessing the external resources provided via links does not play an essential role in students' progress in the non-degree MOOC. Furthermore, students' goals both in MicroMasters and MOOCs have the potential to impact these interactions. For example, research showed that micro-credential programs offered by MOOC providers, like MicroMasters, attract highly educated worker students looking to change career fields, advance skills and knowledge for their current job or even to start their own business (Dillahunt et al., 2016; Ho et al., 2015; Kizilcec & Schneider, 2015; Rivas et al., 2020; Zhenghao et al., 2015). These goals can lead MicroMasters students to seek additional resources to advance their knowledge and practice in various contexts to apply their new skill sets to promote their careers. However, the MOOC course (CS1301) was an introductory programing language course, and it mostly appeals to novice programming students trying to master the basics of the course before seeking additional resources.
- As expected for online courses, loading and playing videos have important roles in the learning outcomes and are found in the list of influential parameters for all considered courses. Studies regarding students' interactions with videos in online learning environments showed an association between students' performance and video loading/playing behaviors (Giannakos et al., 2015; He et al., 2018; Li & Tsai, 2017; Mbouzao et al., 2020). One of these studies, Mbouzao and his colleagues (Mbouzao et al., 2020) predicted students' success in MOOCs based on the video interaction indexes. Researchers reported that students' interaction with the videos at the beginning and midpoint of the course successfully predicted whether students would pass or fail a MOOC course.
- It is noted that, in the regular MOOC, the participants paused videos more often. This activity positively affected their outcome. This could be related to the fact that this non-degree MOOC is designed such that the learners could pause and practice programming skills that are explained in the video lectures but in the MicroMasters courses the practice opportunities embedded within a video lecture are limited. Similar to the regular MOOC is also the CSE6040x that is a programming course in which pausing videos was found among the important activities of learners, though its contribution in prediction is much lower than that of the MOOC. Pausing the video is one of the indicators of

active learning in online environments, and it helps students process information at their pace, thus decreasing the cognitive load (Doolittle et al., 2015). Moreover, pausing, rewinding, and repeatedly watching the video promotes interaction with educational content through video clips so that students can organize and integrate new information with the previous comprehension (Schreiber et al., 2010).

- According to the results, problem interaction has a more significant impact on MicroMaster students than on those who are enrolled in the non-degree MOOC. That could be due to the course requirements as the students in micro-credential degrees are expected to complete most of the assignments and exams, but this is not the case for most of the participants in the non-degree MOOC.
- Displaying video captions and showing corresponding transcripts are key activities for all MicroMasters courses but they are not identified as the top-ranked predictors in the non-degree MOOC. This implies that the students need to pay more attention to the details and get a profound understanding of materials to enhance their learning. This activity enables the learners to go back and check the transcript later once they are reviewing the materials or looking for specific content to solve assignments or take exams.
- The clickstream events "seq_goto" and "seq_next" are among the most important activities of students in the non-degree MOOC. This could be related to the differences in the audiences and their overall goals in registering for these online courses. In the MicroMasters program, students are registered to gain general skills rather than looking for specific topics to learn. Hence, these students need to go over most of the course contents in the provided order and follow the instructor's pace to be able to successfully pass these kinds of degree programs. In contrast, learners of non-degree MOOCs can navigate through the course contents and access particular topics based on their needs since the participants of such courses are often seeking customized content and not necessarily want to earn certificates. This is in agreement with the findings from previous work by Merzdorf and Douglas (2020) in which they found that independent learners were eager to skip some course contents, while learners who were registered in a formal program were dedicated to the entire contents.
- Some differences are observed among the important clickstream activities of participants of the MicroMasters courses. For example, "play_video" was ranked as the third most important event-type in CSE6040x while "problem_check" was ranked third for the other two courses. The mentioned differences may originate from the way these courses are designed. For instance, since the same event types are used to construct the predictive models, the lower prediction accuracy of MGT6203x compared to the results obtained for the other two courses indicates that there are additional factors (other than their online engagement with the course contents) that are influencing students to progress. This difference could be due to the more flexible structure of MGT6203x compared to CSE6040x and ISYE6501x. Students could review course materials and perform exercises offline without the need to completely watch the video lectures. In this case, formative assessments are not required throughout the course and only a small number of mandatory homework is assigned to the students; both could decrease the learners' aspiring online engagement. Moreover, problem-check was identified as one of the first five significant factors for ISYE6501x and MGT6203x, however, it was ranked as a later factor for CSE6040x. That could be related to the difference between the assignment design of these courses since CSE6040x assignments are in the form of one weekly notebook programming task rather than the commonly assigned problem-solving homework.
- Our findings provide some noteworthy implications for learning and instructional practices for the MicroMasters program. This study allowed us to capture the learning behaviors of students enrolled in a specialized professional certificate program, with many of whom already holding Bachelor or more advanced degrees. Information about the key clickstream events that our study discovered can be incorporated into instructional practices and may

be beneficial for a wide range of readers. The key implications are summarized for instructors, designers, and researchers.

- Instructors can monitor students' activities that are associated with links and video loading and track those students who fall behind on a regular basis. This would enable instructors to notify students of concern to remind them of their status or to provide them with appropriate resources at the right timing, which could help them stay on track. In addition to focusing on the two clickstream event types that we identified as the most important across all three MicroMasters courses, instructors could take into account other salient types that are specific to each course when designing their course curriculum. If students exhibit a declining trend over time in terms of how actively they access problem sets or any other significant course content, instructors could also adjust the materials to improve the quality of problem sets for future courses.
- Program designers may provide the basis for learning contents to be offered in different modalities (e.g., visual and audio) in an integrated format so that students can spend less time processing the given information. Instructional designers should optimize the time and effort learners need to put into online course content to keep them motivated to stay engaged in the course. Based on our findings, we recommend the MicroMasters course designers consider incorporating strategies to enhance students' access to external web links which offer various related learning materials and resources outside the course platform. Also, designers of courses such as ISYE6501x or MGT6203x may carefully design problem sets and quizzes to increase students' interest in the course content. For example, to sustain students' engagement with the problem sets, it may be helpful to provide students with timely content-related feedback in response to their submissions of answers to given problems. Additionally, students in the MicroMasters courses may benefit from using a tool that enables them to track their pace and progress of completing required course modules, compared to students in the introductory computer programming language MOOC.
- We implemented a variety of ML techniques for the prediction analysis and significant features identification. For the entire analysis, the nonlinear algorithms provided better overall performance than the linear alternatives, which indicates the nonlinear relationship between the examined clickstream events and the learning outcome. Thus, where high accuracy is desired in this type of research, we suggest the nonlinear algorithms (e.g., Random Forest) as the best candidates. The nonlinear ML algorithms typically require more computational time and power. For large clickstream data, we suggest tuning and optimizing the associated parameters in the ML algorithms (e.g., number of input parameters, regularization factor, and hyperparameters) to reduce computational cost while maintaining acceptable prediction accuracy. Although the nonlinear approaches outperformed the linear models, there are situations in which linear models may be preferable since they would be easier to compute and interpret. The applied linear models in this study still provide comparable prediction accuracy and reasonable cross-validation error to those estimated by the nonlinear approaches and thus could be suitable alternatives in such cases.

Conclusion

MicroMasters degree programs provide learners with opportunities to develop in-demand skills that are not only relevant to their specific career fields but are also validated by industry partners and employers. Besides, these certificates count for credit in participating in relevant Master's degree programs in selected institutions. However, various aspects of this type of program including the trends in learners' online engagement and their differences with non-degree individual MOOCs have not yet been sufficiently investigated. To address this research gap, this study sought to investigate the statistical properties of the learners who are registered for a Master's level MicroMasters program in Analytics to better understand the characteristics of this population. Additionally, this study aimed to understand the relationship between online course

activities and learning progress among students enrolled in the graduate-level MicroMasters certificate program in Analytics offered in an affordable MOOC-based learning platform. To this end, large-scale clickstream data was used, and well-established machine learning approaches were applied. These approaches include algorithms to form predictive models along with feature selection techniques. We sought to find predictive models with an optimal number of predictors mainly to reduce model complexity and improve prediction performance. Moreover, using the clickstream database, one of the main objectives of this research was to explore the differences between the online engagement patterns of learners in the MicroMasters program compared to a non-degree individual MOOC.

Based on the findings, we provided implications for researchers as well as the instructors and designers of similar programs. While it may not be feasible to apply traditional methodology such as surveys to all of the students in our study, future studies could combine clickstream data with ML-based measures of students' psychological factors (Xing et al., 2019). In addition to personal factors, one's learnings community and environment can play a crucial role in predicting learning outcomes. Additionally, in the future, this study can be extended by integrating our predictive modeling with other learning analytics techniques such as time series analysis.

From the perspective of the statistical properties of our predictive modeling, our findings revealed that Random Forest outperformed the other applied statistical approaches in predicting learners' rate of progress based on the clickstream event data. Furthermore, we detected "link_clicked", "load_video", "problem_check", and "closed_captions_show" as the most influential predictors of learner progress in the studied program. In addition to analyzing the courses individually, classification algorithms were implemented to determine the average progress rate of learners in the program (higher/lower than the average). In line with the previous remarks, it was observed that more successful students with a percentage of progress above 50% were on average more active in video-related events such as "play_video" and "seek_video".

The difference between the demographic characteristics and motivation of students from the MicroMasters program and those from the non-degree MOOC revealed some significant differences in terms of their behavioral engagement with the online course contents. In the Master's level credential program, successful learners were likely to pay attention to the learning content thoroughly or on a comprehensive level and seek external resources that were relevant to improving knowledge and skills in their professional field. For highly achieving learners from the introductory-level computing MOOC, it seemed crucial to be able to efficiently orient their cognitive resources to appropriate course content and then apply their newly acquired knowledge through videos to solve actual problems in real-time.

Our results affirmed that, for all investigated courses, providing high-quality and fortified video lectures is an essential key for implementing successful online learning irrespective of offering courses in non-degree or micro-credential formats. According to exploring the pattern of online engagement of learners, one notable observation is that provided links and problems indicated high weights in the progress rate of learners enrolled in the MicroMasters program. Thereby, these two course features should be efficiently optimized in the instructional design process.

Since the findings revealed that the participants of the MicroMasters program greatly relied on the video captions and transcript, and with respect to the fact that navigating captions is a time-consuming task, we suggest examining alternative approaches to eliminate the need for frequently referring to the captions. Potential adjustments to the lectures include trying a slower pace to let the learners absorb the content, providing adequate examples in the presentation to enhance their understanding of complicated concepts, and focusing on key topics that are examined later during the problem assignments. In this way, similar to the studied non-degree MOOC, students can pause the video and self-practice the exercises that eventually improve their learning experience and attention. Furthermore, real-time practices are encouraged to be included in the video lectures which could facilitate students learning and eliminate frequently checking the transcripts and searching for doubtful content.

Notes on contributors

Farahnaz Soleimani is a Research Scientist at the Center for 21st Century Universities (C21U) with an interest in interdisciplinary fields. She earned her Ph.D. and Master's degree in Civil Engineering from the Georgia Institute of Technology with a focus on applied data science and statistics. Her emphasis is on the application of machine learning techniques, data analysis, and learning analytics.

Jeonghyun Lee is a Director of Research in Education Innovation at Georgia Institute of Technology's Center for 21st Century Universities (C21U). Based on her background as a learning scientist, she is interested in how the application of technology facilitates effective learning as well as student success in higher education settings. She earned her Ph.D. degree in educational psychology.

Meryem Yilmaz Soylu is a Research Scientist at the Center for 21st Century Universities (C21U). She holds a Ph.D. in Educational Psychology and a master's degree in Instructional Technology. Her main research interest lies in investigating the impact of self-regulation and motivational processes in writing, learning, and video gaming within various media, including online, blended, and face-to-face learning environments.

ORCID

Farahnaz Soleimani http://orcid.org/0000-0001-7809-2978
Jeonghyun Lee http://orcid.org/0000-0003-1497-0561
Meryem Yilmaz Soylu http://orcid.org/0000-0003-3080-4686

References

Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J., & Radi, N. (2017, May). *Machine learning approaches to predict learning outcomes in massive open online courses* [Paper presentation]. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 713–720). IEEE. https://doi.org/10.1109/IJCNN.2017.7965922

Cabral, P. B., Jorge, N. R., & van Valkenburg, W. (2017). *MicroMasters: the pursuit of the holy grail in online learning* [Paper presentation]. In Proceedings of the Online, Open and Flexible Higher Education Conference: Higher Education for the Future (pp. 219–225).

Dalipi, F., Imran, A., & Kastrati, Z. (2018, April). *MOOC dropout prediction using machine learning techniques: Review and research challenges* [Paper presentation]. In 2018 IEEE Global Engineering Education Conference (EDUCON) (pp. 1007–1014). IEEE. http://doi.org/10.1109/EDUCON.2018.8363340

Davis, D., Kizilcec, R., Hauff, C., & Houben, G. J. (2018). Scaling effective learning strategies: Retrieval practice and long-term knowledge retention in moocs. *Journal of Learning Analytics*, *5*(3), 21–41. https://doi.org/10.18608/jla.2018.53.3

De la Roca, M., Morales, M., Teixeira, A., Hernández, R., & Amado-Salvatierra, H. (2018, September). *The experience of designing and developing an edX's micromasters program to develop or reinforce the digital competence on teachers* [Paper presentation]. In 2018 Learning with MOOCS (LWMOOCS) (pp. 34–38). IEEE. https://doi.org/10.1109/LWMOOCS.2018.8534697

Deng, R., Benckendorff, P., & Gannaway, D. (2020). Learner engagement in moocs: Scale development and validation. *British Journal of Educational Technology*, *51*(1), 245–262. https://doi.org/10.1111/bjet.12810

Dillahunt, T. R., Ng, S., Fiesta, M., & Wang, Z. (2016). *Do massive open online course platforms support employability?* [Paper presentation]. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, 233–244. ACM. https://doi.org/10.1145/2818048.2819924

Doolittle, P. E., Bryant, L. H., & Chittum, J. R. (2015). Effects of degree of segmentation and learner disposition on multimedia learning. *British Journal of Educational Technology*, *46*(6), 1333–1343. https://doi.org/10.1111/bjet.12203

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Giannakos, M. N., Chorianopoulos, K., & Chrisochoides, N. (2015). Making sense of video analytics: Lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course. *The International Review of Research in Open and Distributed Learning*, *16*(1), 260–283. https://doi.org/10.19173/irrodl.v16i1.1976

Gregori, E. B., Zhang, J., Galván-Fernández, C., & de Asís Fernández-Navarro, F. (2018). Learner support in moocs: Identifying variables linked to completion. *Computers & Education*, *122*, 153–168. https://doi.org/10.1016/j.compedu.2018.03.014

He, H., Zheng, Q., Dong, B., & Yu, H. (2018). *Measuring student's utilization of video resources and its effect on academic performance* [Paper presentation]. 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), 196–198. IEEE. https://doi.org/10.1109/ICALT.2018.00053

Ho, A., Chuang, I., Reich, J., Coleman, C., Whitehill, J., Northcutt, C., … Petersen, R. (2015). Harvardx and mitx: Two years of open online courses fall 2012-summer 2014. Available at SSRN 2586847. https://doi.org/10.2139/ssrn.2586847

Hollands, F., & Kazi, A. (2019). *Benefits and costs of mooc-based alternative credentials: 2018-2019 results from end-of-program surveys*. Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

Ingolfsdottir, K. (2016). Winds of change in higher education. *Trends in Pharmacological Sciences*, *37*(12), 990–992. https://doi.org/10.1016/j.tips.2016.09.008

Kizilcec, R. F., & Schneider, E. (2015). Motivation as a lens to understand online learners: Toward data-driven design with the OLEI scale. *ACM Transactions on Computer-Human Interaction*, *22*(2), 1–24. https://doi.org/10.1145/2699735

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014, October). *Predicting mooc dropout over weeks using machine learning methods* [Paper presentation]. In Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs (pp. 60–65). https://doi.org/10.3115/v1/W14-4111

Lee, H., & Kizilcec, R. F. (2020). Evaluation of fairness trade-offs in predicting student success. arXiv preprint arXiv:2007.00088. https://doi.org/10.48550/arXiv.2007.00088

Li, Q., Baker, R., & Warschauer, M. (2020). Using clickstream data to measure, understand, and support self-regulated learning in online courses. *The Internet and Higher Education*, *45*, 100727. https://doi.org/10.1016/j.iheduc.2020.100727

Li, L. Y., & Tsai, C. C. (2017). Accessing online learning material: quantitative behavior patterns and their effects on motivation and learning performance. *Computers & Education*, *114*, 286–297. https://doi.org/10.1016/j.compedu.2017.07.007

Littenberg-Tobias, J., & Reich, J. (2020). Evaluating access, quality, and equity in online learning: A case study of a mooc-based blended professional degree program. *The Internet and Higher Education*, *47*, 100759. https://doi.org/10.1016/j.iheduc.2020.100759

Liang, J., Li, C., & Zheng, L. (2016, August). *Machine learning application in MOOCs: Dropout prediction* [Paper presentation]. In 2016 11th International Conference on Computer Science & Education (ICCSE) (pp. 52–57). IEEE. https://doi.org/10.1109/ICCSE.2016.7581554

Makhlouf, J., & Mine, T. (2020). Analysis of click-stream data to predict stem careers from student usage of an intelligent tutoring system. *Journal of Educational Data Mining*, *12*(2), 1–18. https://doi.org/10.5281/zenodo.4008050

Mbouzao, B., Desmarais, M. C., & Shrier, I. (2020). Early prediction of success in mooc from video interaction features. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial intelligence in education. AIED 2020. Lecture notes in computer science* (vol. *12164*). Springer. https://doi.org/10.1007/978-3-030-52240-7_35

Merzdorf, H. E., & Douglas, K. A. (2020). *Surveying motivation and learning outcomes of advanced learners in online engineering graduate moocs* [Paper presentation]. 2020 IEEE Frontiers in Education Conference (FIE), October, 1–4. IEEE. https://doi.org/10.1109/FIE44824.2020.9273965

Moreno, A. A. (2019). *Predicting student dropout in a micromasters program* [Doctoral dissertation]. Massachusetts Institute of Technology. https://hdl.handle.net/1721.1/122250

Ponce-Cueto, E., & Caplice, C. (2019). *MicroMasters are not just a bunch of MOOCs: Lessons from the first MicroMasters program* [Paper presentation]. In Proceedings of EMOOCs 2019: Work in Progress Papers of the Research, Experience and Business Tracks.

Rivas, M. J., Baker, R. B., & Evans, B. J. (2020). Do moocs make you more marketable? An experimental analysis of the value of moocs relative to traditional credentials and experience. *AERA Open*, *6*(4), 233285842097357. https://doi.org/10.1177/2332858420973577

Rizvi, S., Rienties, B., Rogaten, J., & Kizilcec, R. F. (2020). Investigating variation in learning processes in a futurelearn mooc. *Journal of Computing in Higher Education*, *32*(1), 162–181. https://doi.org/10.1007/s12528-019-09231-0

Schreiber, B. E., Fukuta, J., & Gordon, F. (2010). Live lecture versus video podcast in undergraduate medical education: A randomized controlled trial. *BMC Medical Education*, *10*(1), 68. http://www.biomedcentral.com/1472-6920/10/68 https://doi.org/10.1186/1472-6920-10-68

Stöhr, C., Stathakarou, N., Mueller, F., Nifakos, S., & McGrath, C. (2019). Videos as learning objects in moocs: A study of specialist and non-specialist participants' video activity in moocs. *British Journal of Educational Technology*, *50*(1), 166–176. https://doi.org/10.1111/bjet.12623

Sweeney, M., Lester, J., Rangwala, H., & Johri, A. (2016). Next-term student performance prediction: A recommender systems approach. *Journal of Educational Data Mining*, *8*(1), 22–51. https://doi.org/10.5281/zenodo.3554603

Xing, W., Tang, H., & Pei, B. (2019). Beyond positive and negative emotions: Looking into the role of achievement emotions in discussion forums of moocs. *The Internet and Higher Education*, *43*, 100690. https://doi.org/10.1016/j.iheduc.2019.100690

Zhenghao, C., Alcorn, B., Christensen, G., Eriksson, N., Koller, D., & Emanuel, E. J. (2015). Who's benefiting from MOOCs, and why. *Harvard Business Review*, *25*(1), 2–8. https://hbr.org/2015/09/whos-benefiting-from-moocs-and-why