Journal of Research in Applied Linguistics

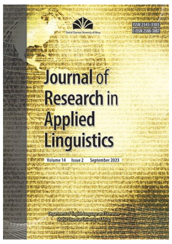
ISSN: 2345-3303 – E-ISSN: 2588-3887 – https://rals.scu.ac.ir

Published by Shahid Chamran University of Ahvaz

Please cite this paper as follows:

del Peral Pérez, J. J. & Castrillo de Larreta-Azelain, M. D. (2023). Empowering teachers in LMOOC design by using a taxonomy of participants' temporal patterns. *Journal of Research in Applied Linguistics*, *14*(2), 116-134. https://doi.org/10.22055/RALS.2023.43579.3048

Shahid Chamran University of Ahvaz



Research Paper

Empowering Teachers in LMOOC Design by Using a Taxonomy of Participants' Temporal Patterns

Juan José del Peral Pérez[1] & María Dolores Castrillo de Larreta-Azelain[2]

[1] Corresponding author, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain; *jperper220y@g.educaand.es*

[2]Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain; *mcastrillo@flog.uned.es*

Received: 28/04/2023 Accepted: 17/09/2023

Abstract

A decade of research into MOOCs (massive online open courses) for language learning (LMOOCs) shows that they seem to have consolidated their position as a subfield of computer-assisted language learning (CALL). Since the appearance of LMOOCs in 2013, 3 key systematic reviews have been carried out; these confirm that research into student profiles is a recurring trend, with the focus on avoiding dropout rates by creating personalized learning pathways. One of the challenges for teachers and LMOOC developers is that they are not cognizant of their students or their study habits. If we could learn how students organize their study in LMOOCs, a taxonomy could be established according to their profiles. This would enable teachers and LMOOC developers to improve their course design and so create personalized learning pathways, making the courses better suited to students' specific learning preferences. In this study, we use techniques of learning analytics (LA) to explore the temporal patterns of LMOOC participants in order to understand the way they manage and invest their time during their online courses. As a result of this study, we propose a new taxonomy of LMOOC participant profiles based on temporal patterns—one which would provide teachers with a tool to support them when personalizing the design and development of LMOOCs and which would, therefore, help them adapt their courses to the specific learning preferences of each profile.

Keywords: LMOOCs; Temporal Access Patterns; Learning Analytics (LA); Participants Profiles; LMOOC Teachers; Learning Pathways.

1. Introduction

After a decade of research into MOOCs (massive online open courses) for language learning (LMOOCs)—first defined as "dedicated Web-based online courses for second languages with unrestricted access and potentially unlimited participation" (Barcena & Martín-Monje, 2014, p. 1)—, they are acknowledged to still be an emerging field (Chong et al., 2022; Panagiotidis, 2019). Furthermore, scholars consider that LMOOCs have consolidated their position as a subfield of computer-assisted language learning (CALL) (Díez-Arcón & Martín Monje, 2023; Gillespie, 2020; Martín-Monje & Borthwick, 2021), as evidenced both by the significant growth in the number of papers on the subject (Sallam et al, 2022; Zeng et al., 2020) and the fact that a special interest group (SIG) on LMOOCs was launched in 2017 within the EUROCALL (European Association for Computer Assisted Language Learning) Association.

In the last decade (2013-2023), five relevant systematic reviews of LMOOCs have been published, all of which including studies and research on the participants' profiles: a recurring research trend in LMOOCs, as will be summarized in the following paragraphs. A growing number of scholars are exploring the identification of user profiles with two main objectives: firstly, to avoid the endemic low completion rates (e.g., Hsu, L., 2023; Jitpaisarnwattana et al., 2021; Jordan, 2015; Liyanagunawardena et al., 2013; Martín-Monje et al., 2018), which in MOOCs are estimated to be close to 90% (Poy & Gonzales-Aguilar, 2014; Read & Sedano, 2021; Reich, & Ruipérez-Valiente, 2019), and secondly to improve the learning experience by designing personalized learning paths adapted to the different needs of the participants, thus diminishing the likelihood of teaching strangers (Chong et al., 2022; Kim & Chung, 2015; Zhang et al., 2023)

2. Literature Review

The first systematic MOOC review by Liyanagunawardena et al. (2013), after analyzing forty-five peer reviewed papers, categorizes the literature into eight different areas of interest, one of which is related to participant profiles: "6

participant focused: considering aspects related to the learners participating in MOOCs" (p. 212). This work is considered a standard reference in the field (Díez Arcón et al. 2022, p. 3) given that it "… finds its correspondence with the CALL disciplinary standards for qualitative research synthesis (QRS)." It states that in this initial phase of the research on MOOCs, most of the studies mainly present demographics regarding the characteristics of the participants' profiles. Furthermore, despite acknowledging the high dropout rates in these courses, they claim not to have access to data on MOOC completion rates. Exploration into the strategies used by active MOOC participants is, thus, encouraged, as this could help reduce low completion rates by providing insight into possible solutions to information overload in MOOC environments.

The systematic LMOOCs review by Sallam et al. (2022), which covers the period between 2012 and 2018, also concluded that research focusing on the profile of the participants is, with a percentage of 21,6%, one of the most popular categories among of the 149 studies (71 qualitative and 78 quantitative) analyzed. In this review, the eight categories established by Liyanagunawardena et al. (2013) were employed to identify a common underlying aim of these papers: the attempt to understand and thus reduce the low completion rates of these courses by drawing information from participant profiles. Díez Arcón and Martín-Monje (2023) replicate the methodology by Liyanagunawardena et al. (2013), in their systematic review of research on LMOOCs published between 2019 and 2021, 29% of the 87 papers included in their study, a significant increase compared to the previous review (Sallam et al., 2022), focus on participant profiling, thus consolidating this type of study as a substantial trend in LMOOC research. In their investigation of the instructional and assessment features of 100 LMOOCs, Chong et al. (2022) also affirm that studies focusing on participants are a major trend in LMOOC research. After identifying the need for systematic analysis and review concerning these issues, they advocate for the provision of multiple learning pathways in order to cater to the individual differences of participants. In the latest paper to date on LMOOCs, Zhang and Sun (2023) use a text-mining approach based on machine-learning techniques to review 71 high quality LMOOC studies that were published up to 2021. As a result, they identify nine research-topics, including "learners and learning in LMOOCs" and state that "… the most promising future direction in T8 research involves learner behavior" (Zhang et al., 2023, p. 85)

The abovementioned studies show that understanding how LMOOC participants relate to the different elements of the course are of great help to LMOOC developers and instructors in designing courses that are tailored to learners' needs, characteristics, and objectives. Furthermore, LMOOC learner profiling allows researchers to subsume participants under fractions of a whole group, thus preventing the identification of individual learning progress, an unaffordable task in massive online educational settings of thousands of learners (Sunar et al., 2020).

2.1. Participant profiles in MOOCs and LMOOCs

Learning analytics (LA), which in recent years has emerged as one of the most widely-cited future trends in education in international reports (Alexander et al., 2019; Brown et al., 2020; Lang et al., 2017; León Urrutia et al., 2017), has become a key tool for improving strategic decisions based on a deeper understanding of student behavior. LA seeks to use data about students and their online contexts to analyses, through main indicators, how the learning process is taking place, and how significant aspects of the educational experience may be improved (Brown et al., 2020; Lester et al., 2018; Ifenthaler et al., 2019; Maseleno et al., 2018; Romero & Ventura, 2020; Sclater, 2017).

Furthermore, the identification of learner profiles is one of the main objectives of LA applied to the field of MOOCs, along with intelligent feedback; the presentation of recommendations; early intervention, or the adaptation and personalization of the learning process. For Chatti et al. (2012), the aim should be the creation of a polycontextual learning profile that allows the elaboration of a detailed picture of the learner's activities in a broader learning context, in addition to the concentration of all the learners' educational actions in the different distributed systems, with the purpose of obtaining more accurate analytical results and more effective proposals for the learners.

Despite the recent increase in the number of studies that use various LA techniques to analyse data generated by MOOCs (Anderson et al., 2014; Coleman et al., 2015; Ferguson & Clow; 2015; Hill, 2013; Ho et al., 2015; Kizilcec et al., 2013; Khalil & Ebner, 2017; Koller et al., 2013; Li & Baker, 2018; Maya-Jariego et al., 2020; Poellhuber et al., 2019; Reich, 2014; Sunar et al., 2020; Xu & Yang, 2016), there are still few studies that apply LA on the specific subtype of LMOOCs. Pioneering research of note has been carried out by Martín-Monje et al. (2018), who analyzed the results of a study that used LA on the LMOOC (i.e., how to succeed in the English B1 level exam), framed in the European project: E-Learning, Communication and Open-Data: Massive Mobile, Ubiquitous and Open Learning (ECO). The dataset generated in the LMOOC was processed using Microsoft Excel software and the SPSS statistical package, after which an inferential statistical analysis of students' online interaction was performed and the profiles and types of learning objects most strongly related to successful course completion were identified. Lee et al. (2018) also employed LA for a descriptive statistics analysis of data collected from the LMOOC Speak English Professionally: In Person, Online, and on the Phone, an LMOOC offered by the Georgia Institute of Technology through the Coursera platform. As in the study by Martín-Monje et al. (2018), the results of this work point to a correlation between video viewing and assignment submission and successful course completion. Castrillo and Sedano (2021) used an LA tool embedded within the MOOC platform and, subsequently, applied statistical methods to the participants' data. The results of their study reflect the impact of an appropriate LMOOC design on the success of the course, which was evidenced by a strong engagement and a high performance. Similarly based on the third edition of an LMOOC, the study conducted by del Peral (2019) employed the LA techniques on logs that recorded all the activity by the course participants. Again, statistical methods were used in this work to identify a series of temporal patterns of platform use by students, as will be analyzed in more detail in the following section.

All the aforementioned authors conclude that the LA techniques allow for inferences about the students' learning process in LMOOCs and for predictions of their behavior, which contributes to the improvement of the learning environment.

In recent years, numerous studies that classify MOOC participants according to common distinguishing features have emerged. Taxonomies of LMOOC participants from scientific literature collected up to 2021 are classified by del Peral (2022) into four broad groups: profiles based on motivation and intention, profiles based on interaction with course elements, profiles based on social interactions, and temporal profiles.

2.2. Exploring the Temporal Dimension in MOOCs and LMOOCs

The description of the quantitative and qualitative degree of learner interaction with certain elements of the MOOC, as presented by most of the taxonomies in the aforementioned studies, provides valuable information on the attachment and engagement of participants with MOOCs and on the type of content preferred. Time is one of the most influential variables on learning success (Chaker, & Bachelet, 2020; Li et al., 2022) and on second language learning performance (iBialystok, 1978; Krashen & Scarcella, 1978; Mehnert, 1998; Tavakoli & Foster, 2011). However, exploration of the temporal patterns of MOOC participants with the aim of understanding the way they manage and invest their time during the online course remains limited (Li et al., 2022); the employment of tools, and the implementation of the LA techniques that allow for the identification of students' behavioral time patterns, remain an important task in educational data mining. Research is significantly lacking on aspects that may reveal how LMOOC participants manage the time they spend on the course. For instance, a user who completes a LMOOC in a single intensive session, doing all the activities and participating in the forums, would be categorized in the same way as a user who schedules the course in shorter, extended sessions, although these profiles clearly exhibit differing characteristics and learning preferences.

Among the few taxonomies that study some aspect of temporal course management are those by Kizilcec et al. (2013), Alario Hoyos et al. (2014), Coleman et al. (2015), Ferguson and Clow (2015), and Poellhuber et al. (2019). These investigations conduct some kind of longitudinal analysis of student engagement during the periods when new content is posted on the MOOC or when an activity is due. Furthermore, in the study by Rodrigues et al. (2016), factors such as the continuity in the completion of activities, or the existence of periods of absence, are taken into account in the definition of their classification, whereas the research by Arora et al. (2017) describes access to MOOCs after their completion as a distinctive feature of explorers.

Several further studies classify MOOC participants according to the way they interact with the course over time. One of these is the prediction model designed by Halawa et al. (2014), which aimed to identify the moment at which

MOOC participants are in danger of abandoning the course. Based on the frequency of interaction with the course, this study used visualization methods to describe four patterns of MOOC learners' persistence:

1. Continuous persistence: Learners who visit the course every few days and who tend to spend several days on each unit.
2. Continuous persistence with prolonged absences: Learners who, at the beginning of the course, tend to interact with it every few days, but are afterwards absent for more than 10 days, on one or more occasions, after which they resume the course.
3. Burst persistence: Students who visit the course occasionally. They usually sample content from different units in every session.
4. Dropouts: Learners who start with a pattern of continuous or burst persistence but vanish before the end of the course.

More recently, Li et al. (2022) applying the LA techniques to data from more than 12,000 participants in a financial MOOC in China, have defined seven different time investment and management patterns in MOOCs learners. They conclude that there is a direct correlation between study time organization by learners and performance and suggest specific cues for the improvement of the learning experience provided by MOOCs.

To the best of our knowledge, the first study to specifically research temporal patterns in an LMOOC by applying the LA and statistical techniques was by del Peral (2019). This study identified five LMOOC user profiles according to their interaction with the course over time, taking into account both the number of sessions and the repetition of time slots by users. After contrasting the fundamental characteristics of each type of student, the author concluded that belonging to any of the different profiles that complete the LMOOC does not imply a higher performance in the course or a better grade (see Table 1):

Table 1. *LMOOC Participant Profiles Based on Temporal Patterns (Adapted From del Peral, 2019)*

Preconsumers	Users who never access the main MOOC content. Three subgroups: "Only enrolled": learners who enroll but never access the course. "Only forums": learners who do not access any of the course modules (neither the initial modules, nor the main modules, nor the assessment tests), but do access the forums. Learners who interact with the course (accessing the initial module and/or taking the pre-test) but do not access any of the three main modules of the MOOCs.
Intensive Users	Users who access the main content of the course through 1 or 2 sessions. These are, as can be deduced, users who spend more time on each session, but less total time on the MOOCs, and access the course forums less frequently.
Routine Users	Users who access the main contents of the course throughout more than 2 sessions and less than 15, always within the same time slot. In addition to obtaining the highest marks among all profiles, these users exhibit higher values of total time, number of sessions and number of accesses to forums than do intensive users, but lower values than flexible and exhaustive users.
Flexible Users	Users who dedicate between 2 and 15 sessions in different time slots to the course. The total time; the number of sessions, and the number of accesses to the forums is higher than in intensive and routine users, but lower than in exhaustive users.
Exhaustive Users	Users who dedicate more than 15 sessions to the course. These are the users who spend the most time on the course and who interact the most with the forums, but who have the shortest sessions.

The general aim of this paper is the identification, within two different LMOOCs, of the taxonomy of profiles based on temporal patterns described by del Peral (2019), so that the traits of each type of user can be considered in the methodological design of LMOOCs and in the adaptation of courses to the specific learning preferences of learners. A second objective of this study, deriving from the first, is the analysis of the possible differences between the different participant profiles according to the characteristics of each LMOOC. One of the LMOOCs employed in this analysis is Inglés Profesional/Professional English (IP hereafter), a course aimed at intermediate English learners whose objective was the improvement of their command of English for working purposes. This course was the same LMOOC for which del Peral's taxonomy was defined, so further the LA techniques will be applied to it. The second LMOOC is Puertas Abiertas (PA hereafter), an LMOOC aimed especially at displaced individuals such as refugees and migrants, whose objective is to help learners use Spanish in common situations, as well as to teach basic aspects of Spanish life and culture. The differences in the type of user of the two LMOOCs suggest differences in the prevalence of the different profiles.

The steps we will follow in the present study are, therefore, as follows:

1. Employ the LA techniques to identify the same profiles of LMOOC participants described by del Peral (2019) in the two LMOOCs.
2. Analyze the data obtained in order to determine firstly, whether the profiles identified in del Peral's taxonomy can be found in IP using a different LA technique and, secondly, whether they can be extrapolated to PA and so be considered common to all LMOOCs.
3. Compare the traits of the different profiles in the two LMOOCs so as to determine whether the inherent characteristics of each course and their participants produce disparities in the prevalence and the level of performance.

The general hypothesis posed in this study is, thus, the following: It is possible to identify the taxonomy of LMOOC user profiles based on temporal patterns of use described by del Peral (2019) employing different LA techniques on the LMOOC IP and on a new LMOOC, PA, but the levels of prevalence and performance will vary depending on the particular characteristics of each LMOOC and its participants.

As a result of the research objective and the hypothesis, this paper seeks to answer the following general and specific research questions (RQ):

- General RQ1: Is it possible to set up a taxonomy of LMOOC participant profiles based on temporal patterns of use?
    - Specific RQ1: How many profiles can be identified?
    - Specific RQ2: What are the identifying features of each profile?
- General RQ2: Are there any differences in the various profiles depending on the LMOOC from which the data are obtained?
    - Specific RQ3: What are the differences regarding:
        - Prevalence of profiles
        - Degree of performance
        - Course access times
        - Number of sessions?
    - Specific RQ4: What inherent features of the LMOOCs design produce such differences?
    - Specific RQ5: What characteristics of the students in the courses produce such differences?

3. Methodology

IP and PA were both LMOOC offered by UNED Abierta through the Open edX platform and were followed by 8,326 and 2,252 participants, respectively. IP was a 3-week LMOOC (running from April 18th to May 9th, 2017), aimed mainly at participants with an intermediate English level and at first-year students of a Tourism degree at UNED, who were rewarded with 1 ECTS credit after its completion. The course started with a test to assess the B1 MCERL English level required for success in the course. IP was divided into three different modules, each of which included four different videos that portrayed the life of a British worker at an American multinational. The main activities in the LMOOC were multiple-option tests based on these videos and on audio recordings, as well as a final test. In each module, there was a forum in which facilitators encouraged interaction by highlighting grammar components of the content. Finally, there was a general forum for technical and global issues.

PA, which lasted for 6 weeks (from January 15th to March 10th, 2019), was designed and developed by the UNED ATLAS research group within the framework of the European project MOONLITE (massive open online courses enhancing linguistic and transversal skills for social inclusion and employability). The openness in this LMOOC was extended to the completion certificate, which was issued free of charge, something uncommon among MOOCs. The main objective of PA was the enhancement of participants' social inclusion and employability through the development of linguistic, intercultural, and transversal skills. The four IP modules made use of texts, images, infographics, and audio/video files to introduce the language resources needed to communicate in Spanish in common daily tasks such as getting around the city, making plans, looking for housing, filling out paperwork, going to the doctor or defending one's rights. Performance was assessed through a variety of activities, such as self-corrected multiple-option tests or self-assessment and P2P activities. Social interaction was also promoted via the six course forums: one for general questions, one for technical issues, and one for each of the modules, in which participants shared their experiences and viewpoints and explained the contrast between the daily aspects of their culture and those of their host country (Castrillo & Sedano, 2021; Sanz Gil, 2021).

EdX, the platform where both LMOOC were hosted, allowed the generation of two different types of files: logs in JSON format, containing a record of all the interactions of the participants with the courses, and a spreadsheet in CSV format, which included key information such as the scores of the students in each activity, their final average score, and whether the student was entitled to receive the completion certificate.

From the academic data gathered in these files, it was necessary to select and obtain a group of variables to which LA techniques, more particularly the k-means algorithm, a clustering algorithm belonging to partition algorithms, be applied (Khalil & Ebner, 2016; Liu, 2006; Romero et al., 2016). The variables that were calculated were (1) the total number of sessions, (2) the final grade in the course, and (3) the time distribution of the sessions.

In order to extract these variables, the data in the files were preprocessed using a Java computer program, specifically designed for this purpose, which performed the following tasks:

- Identification of sessions: To extract each of the participant's sessions from the data stored in the log files, it was necessary to establish the threshold value of a user's session. This research chose a value of 30 min, the most common inactivity threshold among Websites (Ho et al., 2015; Mullaney, 2014). Any activity carried out by a participant 30 min after their last recorded event was, therefore, registered as the beginning of a new session.

The method by which the Java program calculated the sessions was the following:

  - For each user, every recorded event in the logs was stored in a list.
  - All the events in the list were sorted in chronological order.
  - A list of sessions containing start and end times was created for each user.

- Data integration: Once the list of sessions was created, the Java program processed the CSV file and integrated for each participant, the information regarding the scores, and the eligibility for the completion certificate.
- Data reduction: To calculate the time distribution of the sessions, the median absolute deviation or MDA, a robust dispersion measure that presents lower sensitivity to extreme values (Leys et al., 2013), was selected as measure of variation of the starting times of the sessions. For the calculation of the MDA, the starting time in seconds of each session was stored in an ordered list, from which the Java program obtained firstly the median value and secondly the MDA, which was subsequently stored and associated to the corresponding user.
- Data modelling: From the variables calculated during the preprocessing phase, the Java program generated a spreadsheet in CSV format. The structure of the spreadsheet is the following:
    1. The first column contained the name of each participant.
    2. The second column contained the final grade of each participant.

3. The third column contained the MDA of the starting times of each participant's sessions.
4. The fourth column contained the total number of sessions of each participant.

- Data transformation: To minimize bias and prevent certain numerical variables from superseding others in the generation of the clusters, a transformation of data into a common range is required prior to the execution of the k-means algorithm (Singh & Singh, 2020). A scaling operation, that is, the transformation of the data to fit a specific scale, was therefore performed using the statistical program *R*. This operation is frequently used in algorithms based on data clustering, such as support vector machines, KNN, or k-means (Tatman, 2018).

4. Results

4.1. General Research Question 1

The first general research question formulated the existence of a taxonomy of LMOOC participants based on temporal patterns of use. In order to identify the different profiles, the k-means algorithm was applied to the scaled values of the three variables using the software *R*. As can be observed in Figure 1, which represents the clustering obtained for each of the samples, the optimal value for PA is 6 and for IP is 3:

Figure 1. Clusters Obtained by the k-Means Algorithm in R for PA (Left) and IP (Right)

To assess the quality of the clustering obtained, the function gap statistic, a method that calculates an error metric based on the sum of squares within a group (Tibshirani et al., 2001) was executed. Although this function confirms the values of 6 and 3 clusters for PA and IP, respectively, it can be observed in Figure 2 that five clusters would also be an appropriate grouping for IP, as the error metric decreases between three and four clusters, then increases again for five clusters, to decrease once more for six clusters:

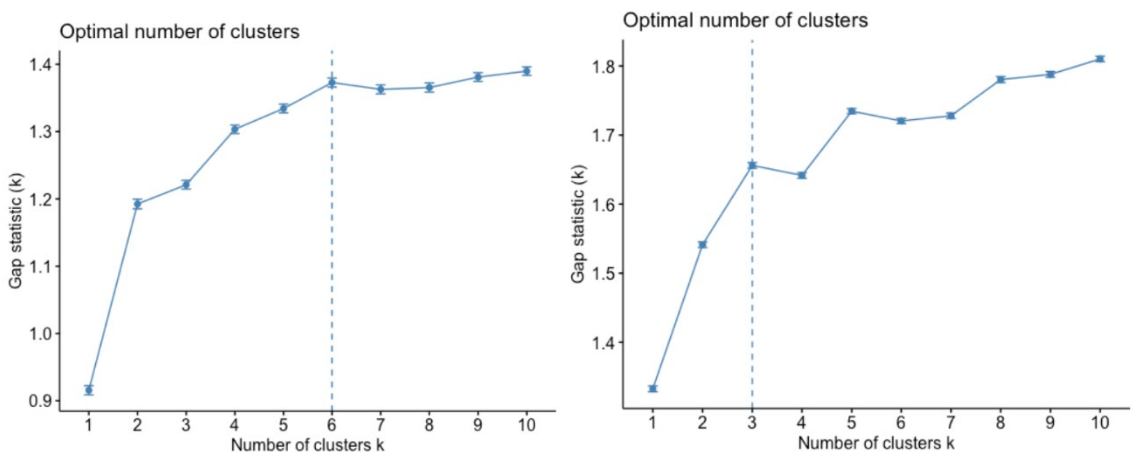


Figure 2. Output of Function Gap Statistics in R for A (Left) and IP (Right)

Figure 3 represents a juxtaposition of the six clusters for PA and the new five-group clustering for IP. In order to determine whether there is an underlying taxonomy common to both courses, a correlation between the profiles identified by the k-means algorithm and the five profiles identified by del Peral (2019) was established. In Figure 3, a coincidence between groups 1 of the two courses can be observed. These groups clearly correspond with the profile of preconsumers, participants with the lowest values for the three variables: the grade, the number of sessions and the concentration of study hours (time dispersion for one-off sessions is 0):

Figure 3. Clusters Obtained by the k-Means Algorithm in R for PA (Left) and IP (Right)

As observed in Figure 3, group 5 of PA and group 4 of IP also show clear similarities in shape, size, position, and distribution of the samples. These two groups present the highest values in the number of sessions, which matches the definition of the exhaustive participants.

An evident likeness in the shape, size, position, and distribution of the samples can also be found between group 3 of PA and group 5 of IP, and between group 4 of PA and group 3 of IP. These four groups exhibit low values of the two variables associated with a good performance in the LMOOC: final marks and number of sessions. In the taxonomy by del Peral (2019), the group characterized by a low number of sessions, no more than two, was the intensive, of whom only 52 out of 8,326 total participants completed the LMOOC successfully. Consequently, as the distinctive characteristic of this group is the low engagement with the course, a renaming as occasional participants would perhaps be more accurate. This group is therefore divided into two different subgroups, as can be inferred from Figure 3: one with little variation in the dispersion of the sessions, named occasional routine participants, and another with a much higher variation in the dispersion of the sessions, named occasional flexible participants.

Finally, Figure 3 shows that the k-means algorithm has divided into two different groups, 2 and 6 of PA, the samples clustered in only one group in IP. However, in view of the shape, size, position, and distribution of these samples, it would be appropriate to conclude that these two groups of PA parallel the hierarchy of the occasional participants and are, in fact, subdivisions of group 2 in IP. Because of the high values in the final marks and the number of sessions of both subgroups, the supergroup has been named constant and the two subgroups routine and flexible.

After this analysis, the first general research question can be affirmatively answered in the affirmative, as the interpretation of the data has led to the definition of a taxonomy based on temporal patterns. From this analysis, the three specific questions will be answered.

4.1.1. Specific RQ1: How Many Profiles Can Be Identified?

The execution of the k-means algorithm points to the existence of four main profiles, two of which are, in turn, divided into two subprofiles:

- Preconsumers
- Exhaustive

- Occasional
    - Routine occasional
    - Flexible occasional
- Constant
    - Routine constant
    - Flexible constant

4.1.2. Specific RQ2: What Are the Identifying Features of Each Profile?

According to the three variables used as the input for the k-means algorithm, the characteristics that define each of the profiles are the following:

- Preconsumers: As can be seen in Tables 2 and 3, preconsumers show minimal levels of activity in the course, with scores close to 0 and average of sessions close to 1.

Table 2. *Preconsumers Statistics for PA*

	Average Mark	Session Dispersion (min)	Number of Sessions
Mean	0.03	8.98	1.53
SD	0.08	17.46	1.18

Table 3. *Preconsumers Statistics for IP*

	Average Mark	Session Dispersion (min)	Number of Sessions
Mean	0.02	8.05	1.34
SD	0.07	17.31	0.72

- Exhaustive participants: Although this group presents the highest values in the number of sessions, as can be seen in Tables 4 and 5, the final marks are surpassed by constant participants, both in PA and in IP. As already anticipated by del Peral (2019), this is a counterintuitive phenomenon because a direct correlation between sessions spent on the course and performance might be expected.

Table 4. *Exhaustive Participants Statistics for PA*

	Average Mark	Session Dispersion (min)	Number of Sessions
Mean	0.87	208.7	40.42
SD	0.25	65	13.06

Table 5. *Exhaustive Participants Statistics for IP*

	Average Mark	Session Dispersion (min)	Number of Sessions
Mean	0.02	158.1	24.26
SD	0.07	69.9	10.58

- Occasional participants: The two subprofiles of this group exhibit poor performance in the course, as seen in Tables 6 - 9. Regarding the dispersion of the sessions, the value varies for each subgroup:
    - Routine occasional participants: The few sessions dedicated to the course tend to be concentrated in the same periods, so the dispersion is low.

Table 6. *Routine Occasional Participants Statistics for PA*

	Average Mark	Session Dispersion (min)	Number of Sessions
Mean	0.11	127.1	5.37
SD	0.14	45.5	3.89

Table 7. *Routine Occasional Participants Statistics for IP*

	Average Mark	Session Dispersion (min)	Number of Sessions
Mean	0.1	124.6	4.17
SD	0.14	50.4	2.51

- Flexible occasional participants: The few sessions of this group tend to be concentrated in different periods, so the dispersion is high.

Table 8. *Flexible Occasional Participants Statistics for PA*

	Average Mark	Session Dispersion (min)	Number of Sessions
Mean	0.14	335.4	5.19
SD	0.22	105.2	4.41

Table 9. *Flexible Occasional Participants Statistics for IP*

	Average Mark	Session Dispersion (min)	Number of Sessions
Mean	0.11	335.8	3.69
SD	0.2	83.8	2.44

- Constant participants: This group exhibits a high number of sessions and the highest marks among all the profiles, as can be seen in Table 10, which integrates for IP the two subprofiles, and Tables 11 and 12 for PA.

Table 10. *Constant Participants Statistics for IP*

	Average Mark	Session Dispersion (min)	Number of Sessions
Mean	0.87	112.3	6.92
SD	0.13	75.2	3.53

- Routine constant participants: The sessions dedicated to the course tend to be concentrated in the same periods, so the dispersion is low.

Table 11. *Routine Constant Participants Statistics for PA*

	Average Mark	Session Dispersion (min)	Number of Sessions
Mean	0.9	72.4	8.28
SD	0.13	44.3	4.2

- Flexible constant participants: The sessions dedicated to the course tend to be concentrated in different periods, so the dispersion is high.

Table 12. *Flexible Constant Participants Statistics for PA*

	Average Mark	Session Dispersion (min)	Number of Sessions
Mean	0.93	200	15.96
SD	0.13	61.6	5.86

4.2. General Research Question 2

The second general research question asked if there were differences in the characteristics of the profiles depending on the LMOOC from which the data were obtained. The answer is affirmative, as it will be made clear by responding to the fourth specific question.

4.2.1. Specific RQ3: Differences

- Prevalence of profiles: As shown in Figure 4, there is a higher proportion of participants belonging to profiles with lower levels of activity and performance in IP. Thus, the rate of preconsumers is much higher in IP (58%) than in PA (43%), as well as in the occasional participants (24% in IP and 18% in PA). On the other hand, in PA a greater number of participants engaged with the course, with a higher rate of exhaustive participants (5% in PA and 2% in IP) and, particularly, of constant participants (35% in PA and 15% in IP).

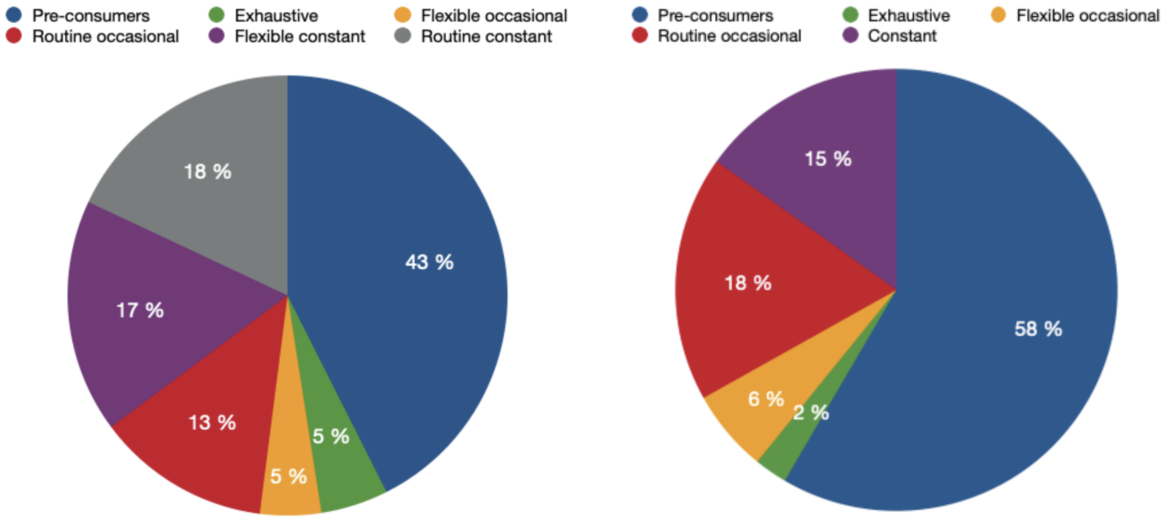


Figure 4. Prevalence of Profiles for PA (Left) and IP (Right)

- Performance: As we can see in Figure 5, the final mark is higher for all PA profiles, particularly for the exhaustive participants, for whom the difference is almost one point out of ten.

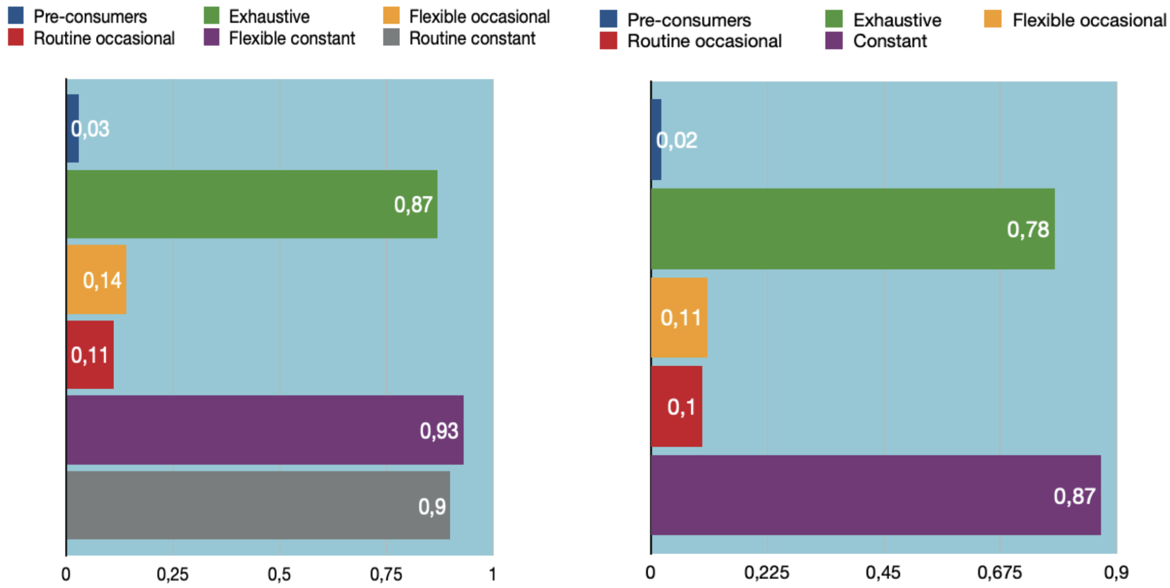


Figure 5. Final Marks for PA (Left) and IP (Right)

- Course access times: As can be seen in Figure 6, the dispersion of the beginning of sessions presents very similar values across all the profiles, except for the exhaustive participants, with 50 more minutes of variation in PA.

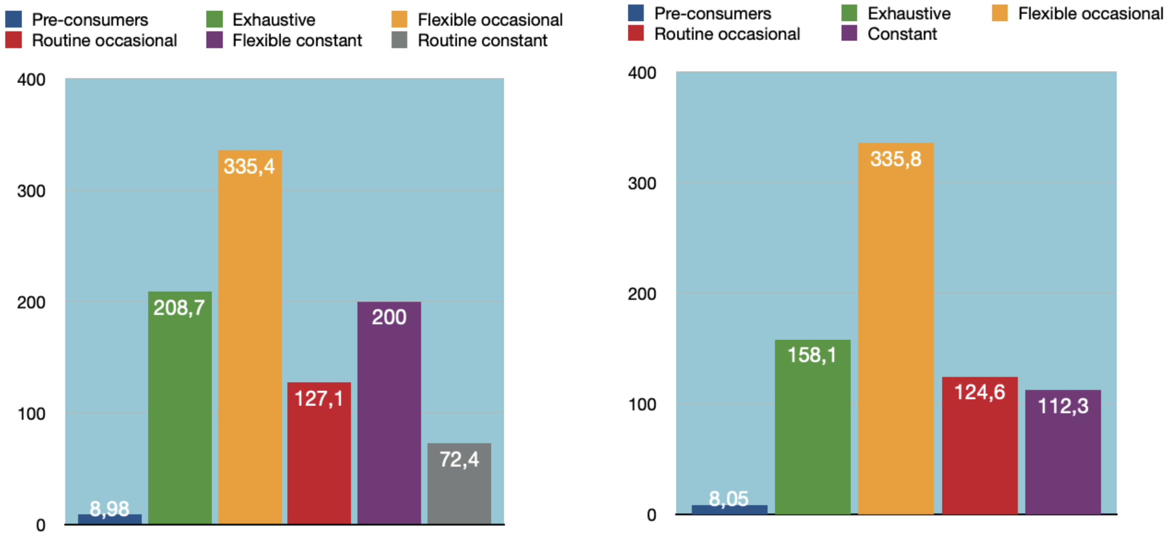


Figure 6. Dispersion of the Beginning of Sessions for PA (Left) and IP (Right)

- Number of sessions: Figure 7 shows that the number of sessions is higher in all PA profiles and that the difference increases in parallel to the degree of involvement of each group, reaching the maximum value in the exhaustive participants, with 16 more sessions in PA than in IP.

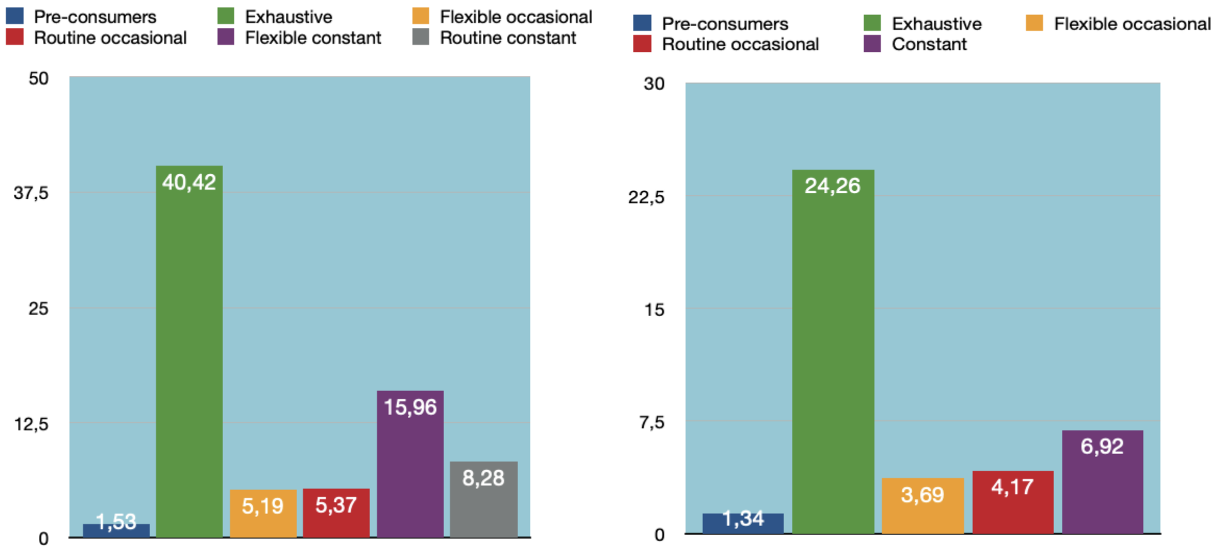


Figure 7. Number of Sessions for PA (Left) and IP (Right)

In conclusion, after the comparison between the two LMOOCs, it can be stated that the only variable for which there are no significant disparities is that of the dispersion of the beginning of the sessions. On the other hand, PA shows higher values in remaining variables throughout all the profiles, as well as a higher rate of participants belonging to groups with greater engagement to the course and a better performance.

4.2.2. Specific RQ4: What Inherent Features of the LMOOCs Design Produce Such Differences?

Although the two courses share several design characteristics, PA was a significantly longer course, with one more module and two more weeks of duration than IP. This is likely to have increased the session number spent on the course. It would not, however, explain the better performance of PA over IP.

4.2.3. Specific RQ5: What Features of the Students in the Courses Produce Such Differences?

The stronger performance of all the PA profiles is noteworthy, especially when the educational level of the participants, which as shown in Figures 8 and 9 is considerably higher in IP, is taken into account. Whereas 71% of IP participants possess graduate or postgraduate studies and only 4% have not completed compulsory secondary education, the corresponding figures for the PA participants are 29% and 37%, respectively:

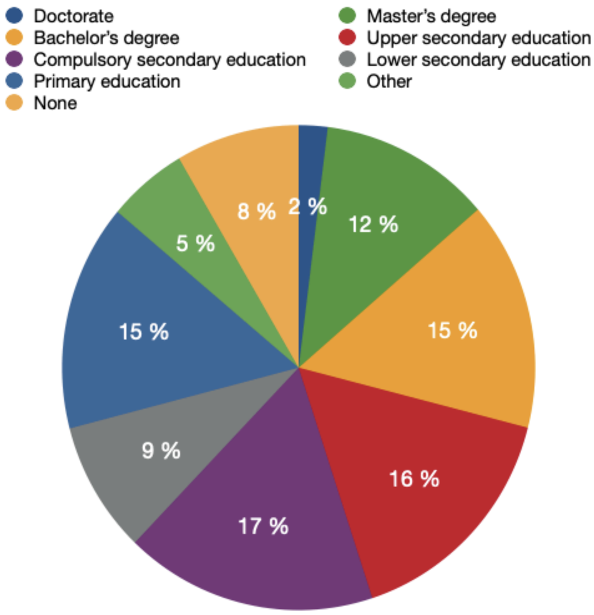


Figure 8. Educational Level of PA Participants

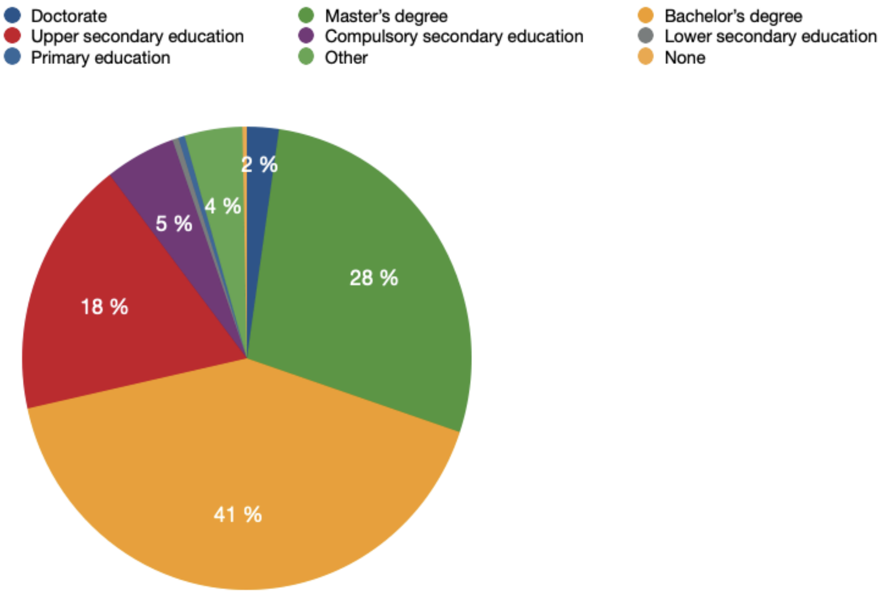


Figure 9. Educational Level of PA IP Participants

These data suggest that the stronger performance of the PA participants may be due to their greater involvement with the LMOOC. This is supported by the characteristics of the target audience of each LMOOC. IP was aimed at professionals and university students, individuals characterized by having limited time to balance the LMOOC with their employment and/or studies. On the other hand, PA was designed for migrants who need to learn Spanish, individuals among whom employment rates tend to be low. This usually translates to more time available to commit to the LMOOC and a stronger motivation to complete it, which would contribute to a better performance on the course.

5. Conclusion

The present study employed LA to describe a 6-group taxonomy of LMOOC profiles based on temporal patterns which can be expected to be found in LMOOCs belonging to the same typology as PA and IP. This statement is based on three facts: Firstly, the distribution and clustering of the samples show a very similar proportion, shape, density, and position for all the groups. Secondly, for both LMOOCs, there exists a clear commonality in the identified profiles, with proportional values in all the variables. Lastly, the fact that the same taxonomy of participants has been identified in two LMOOCs with evident differences in duration, structure, resources, and objective audience suggests that the described profiles are common, at least, to courses of the same typology, that is, LMOOCs with an xMOOC design.

The taxonomy described in this paper can help teachers to adapt LMOOCs to the specific learning preferences of each profile. This would empower teachers and LMOOC developers by enabling them to improve their course design and so create personalized learning pathways, making the courses better suited to students' specific learning preferences (Nurieva & Garaeva, 2020). In this way, teachers can combine the knowledge about the characteristics of the different profiles with initial surveys. This would shed light on the objectives and motivations of students and, particularly, on their time constraints, schedules and study habits. By the combination of this information with early profiling and LA techniques, teachers could provide students with continued individualized help, especially disengaged students, who could benefit from tailored intervention mechanisms. Some of the strategies that teachers and LMOOC developers could implement are gamification, badges, email reminders and notifications, time management guides based on the student's history or modifications to learning paths. By adapting these strategies to the features of the different profiles, teachers would help participants to assimilate the contents of the courses and thus achieve a higher performance, in addition to contributing to the reduction of the dropout rates.

One way in which teachers can integrate the different patterns of time management into the design of LMOOCs would be the inclusion of different activity types. Shorter activities may be particularly appropriate for increasing the motivation of occasional or even exhaustive participants, who tend to dedicate shorter sessions to the course. On the other hand, challenging activities which require greater dedication or present a higher level of difficulty may be suitable when fostering motivation among constant profiles. Teachers could extend this same flexibility to content, particularly to videos, which could be split or shortened based on the different profiles, creating individualized learning routes based on the best suited typology and presentation of activities for each profile (Ahmadi, 2015).

For exhaustive participants, whose greater dedication in time, sessions and activity in the forums does not translate to better grades, teachers could implement reward mechanisms or regular microtasks, so as to boost performance and identify possible difficulties with course content. Additionally, teachers could include this group of students in peer tutoring programs involving participants from low engagement profiles, particularly occasional participants, as this profile of participants would be especially suited for activities which require a greater dedication of time. This type of program has already been implemented in some MOOCs and has contributed to the improvement of completion rates and to student satisfaction levels (Dhorne et al., 2017; Garreta et al., 2015).

Several limitations in this study are worth mentioning. Due to the impossibility of recording some of the student activity outside the platform, the study of downloaded material, for example, the number and duration of some sessions might have been underestimated. In addition, IP included an optional activity based on a podcast application, for which logs were not made available. All this might have resulted in the underestimation of some profiles, likely to be exhaustive and constant participants who may have spent a higher number of sessions on the course; in the case of constant participants, this would have meant transferring them into the exhaustive group.

In addition, it should be remembered that the typology of the two LMOOCs analyzed in this research is xMOOC. LMOOCs belonging to the other predominant typology, c-MOOC, are still scarce and extracting educational data from them is remarkably challenging. Therefore, although this study suggests that the taxonomy described is expected to be inherent to all LMOOCs, perhaps even to all MOOCs regardless of their theme, one future line of enquiry would be the confirmation of this hypothesis resulting from specific research based on data obtained from a language cMOOC.

Other possible lines of research might be the utilization of an alternative clustering algorithm, such as k-medoids or some other partitioning algorithms; methodological triangulation by resorting to other research techniques, such as surveys and interviews, in order to delve into the study routines of the participants and how they impact their performance; or the analysis of additional temporary variables, such as time zones differences or late starts in the course, and their impact on dropout rates.

Finally, in view of the better results exhibited by all the PA profiles despite their lower average level of studies, a further interesting line of research emerges with the potential use of LMOOCs as a methodological complement to secondary education and vocational training, particularly as a measurement of the attention to diversity and as extension material for advantaged or high-capacity students.

Acknowledgments

The empirical data for this paper were collected from the LMOOC (i.e., how to succeed in the English B1 level exam), designed within the EU-funded ECO Project (E-learning, Communication, and Open-data Massive Mobile, Ubiquitous and Open Learning, grant no. 21127) and the LMOOC (i.e., Puertas Abiertas), designed within the European Erasmus+ Program research project MOONLITE (Massive Open Online Courses enhancing linguistic and transversal skills for social inclusion and employability, ref.no.: 2016-1-ES01-KA203-025731). The research presented was undertaken with the support of Universidad Nacional de Educación a Distancia (UNED).

Conflict of Interest

The authors declare that there is no conflict of interest.

References

Ahmadi, N. (2015). Which learning style do you prefer to improve EFL learning? *Journal of Research in Applied Linguistics, 5*(Special Issue), 26-34. https://doi.org/10.22055/rals.2015.11266

Alario-Hoyos, C., Pérez-Sanagustín, M., Delgado-Kloos, C., Parada, H. A., & MuñozOrganero, M. (2014). Delving into participants' profiles and use of social tools in MOOC. *IEEE Transactions on Learning Technologies, 7*(3), 260-266. https://doi.org /10.1109/TLT.2014.2311807

Alexander, B., Ashford-Rowe, K., Barajas-Murphy, N., Dobbin, G., Knott, J., McCormack, M., Pomerantz, J., Seilhamer, R., & Weber, N. (2019). *EDUCAUSE horizon report: 2019 higher education edition*. EDUCAUSE.

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). Engaging with massive online courses. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 687-698). ACM. https://doi.org/10.1145/2566486.2568042.

Arora, S., Goel, M., Sabitha, A. S., & Mehrotra, D. (2017). Learner groups in massive open online courses. *American Journal of Distance Education, 31*(2), 80-97. https://doi.org/10.1080/08923647.2017.1300461

Barcena, E., & Martín-Monje, E. (2014). Language MOOCs: Providing learning, transcending boundaries. In E. Martín-Monje & E. Barcena (Eds.), *Language MOOCs: Providing learning, transcending boundaries* (pp. 175-189). Walter de Gruyter GmbH.

Bialystok, E. (1978). A theoretical model of second language learning 1. *Language learning, 28*(1), 69-83.

Brown, M., McCormack, M., Reeves, J., Brooks, D. C., Grajek, S., Alexander, B., Bali, M., Bulger, S., Dark, S., Engelbert, N., Gannon, K., Gauthier, A., Gibson, D., Gibson, R., Lundin, B., Veletsianos, G., & Weber, N. (2020). *EDUCAUSE horizon report, teaching and learning edition*. EDUCAUSE.

Castrillo, M. D., & Sedano, B. (2021). Joining Forces toward social inclusion: Language MOOC design for refugees and migrants through the lens of maker culture. *CALICO Journal, 38*(1), 79-102. https://doi.org/10.1558/cj.40900

Chaker, R., & Bachelet, R. (2020). Internationalizing professional development: Using educational data mining to analyze learners' performance and dropouts in a French MOOC. *The International Review of Research in Open and Distributed Learning, 21*(4), 199-221. https://doi.org/10.19173/irrodl.v21i4.4787

Chatti, M. A., Dyckhoff, A.L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning, 4*(5 & 6), 318-331. https://doi.org/10.1504/IJTEL.2012.051815

Chong, S. W., Khan, M. A., & Reinders, H. (2022). A critical review of design features of LMOOCs. *Computer Assisted Language Learning*, 1-21. https://doi.org/10.1080/09588221.2022.2038632

Coleman, C. A., Seaton, D. T., & Chuang, I. (2015). Probabilistic use cases: Discovering behavioral patterns for predicting certification. In *2nd ACM Conference on Learning@ Scale* (pp 141–148). ACM.

Díez-Arcón, P., & Martín-Monje, E. (2023). The coming of age of LMOOC research. A systematic review (2019-21). *Innovation in Language Learning and Teaching*, *17*(3), 535-551.

Dhorne, L., Deflandre, J., Bernaert, O., Bianchi, S., & Thirouard, M. (2017). Mentoring learners in MOOCs: A new way to improve completion rates? In *Digital Education: Out to the World and Back to the Campus* (Vol. 10254, Lecture Notes in Computer Science, pp. 29-37). Cham: Springer International Publishing.

Ferguson, R., & Clow, D. (2015). Examining engagement: Analyzing learner subpopulations in massive open online courses (MOOCs). *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (pp. 51-58). Vancouver, Canada: ACM.

Garreta, M., Aguado, G., Mor, Y., Fernandez, C., & Riviou, K. (2015). A peer-mentoring approach for the continuous professional development of educators in a MOOC setting. In *Proceedings of the European Distance and E-Learning Network 2015 Annual Conference* (pp. 149-158). Barcelona, Spain.

Gillespie, J. (2020). CALL research: Where are we now? *ReCALL, 32*(2), 127-144. https://doi.org/10.1017/S0958344020000051

Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. In U. Cress & C. Delgado Kloos (Eds.), *Proceedings of the European MOOC stakeholder summit* (pp. 58-65). P.A.U. Education.

Hill, P. (2013). *Emerging student patterns in MOOCs: A (revised) graphical view*. Retrieved from http://mfeldstein.com/emerging-student-patterns-in-moocs-a-revised-graphical-view

Ho, A. D., Chuang, I., Reich, J., Coleman, C. A., Whitehill, J., Northcutt, C. G., et al. (2015). *Harvardx and mitx: Two years of open online courses fall 2012-summer 2014.* https://doi.org/10.2139/ssrn.2586847

Hsu, L. (2023). What makes good LMOOCs for EFL learners? Learners' personal characteristics and information system success model. *Computer Assisted Language Learning*, *36*(1-2), 1-25.

Ifenthaler, D., Mah, D., & Yau, Y. J. (Eds.). (2019). *Utilizing learning analytics to support study success*. Springer.

Jitpaisarnwattana, N., Reinders, H., & Darasawang, P. (2021) Understanding the roles of personalization and social learning in a language MOOC through learning analytics. *Online Learning, 25*(4), 324-343.

Jordan, K. (2015). Massive open online course completion rates revisited: Assessment, length, and attrition. *The International Review of Research in Open and Distributed Learning, 16*(3), 30-42. http://dx.doi.org/10.19173/irrodl.v16i3.2112

Khalil, M., & Ebner, M. (2015). Learning analytics: Principles and constraints. In S. Carliner, C. Fulford, & N. Ostashewski (Eds.), *Proceedings of EdMedia 2015—World Conference on Educational Media and Technology* (pp. 1789-1799). Montreal, Quebec, Canada: Association for the Advancement of Computing in Education (AACE).

Khalil, M., & Ebner, M. (2016). What is learning analytics about? A survey of different methods used in 2013-2015. In *Proceedings of Smart Learning Conference* (pp. 294-304). Dubai: HBMSU Publishing House.

Kim, P., & Chung, C. (2015). Creating a temporary spontaneous mini-ecosystem through a MOOC. In C. J. Bonk, M. M. Lee, T. C. Reeves, & T. H. Reynolds (Eds.), *MOOCs and open education around the world* (pp. 157-168). New York, NY: Routledge.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *3rd International Conference on Learning Analytics and Knowledge* (pp. 170-179). New York: ACM.

Koller, D., Ng, A., Do, C., & Chen, Z. (2013). Retention and intention in massive open online courses: In depth. *Educause Review, 48*(3), 62-63.

Krashen, S., & Scarcella, R. (1978). On routines and patterns in language acquisition and performance *Language Learning, 28*(2), 283-300.

Lang, C., Siemens, G., Wise, A., & Gaševic, D. (Eds.). (2017). *Handbook of learning analytics*. SOLAR.

Leys, C., Christophe Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764-766. http://dx.doi.org/10.1016/j.jesp.2013.03.013

Li, Q, & Baker, R. (2018). The different relationships between engagement and outcomes across participant subgroups in Massive Open Online Courses. *Computers & Education, 127*, 41-65. https://doi.org/10.1016/j.compedu.2018.08.005

Li, S., Wang, S., Du, J., Pei, Y., & Shen, X. (2022). MOOC learners' time—investment patterns and temporal—learning characteristics. *Journal of Computer Assisted Learning, 38*(1), 152-166. https://doi.org/10.1111/jcal.12597

Liu, B. (2006). *Web data mining*. Springer.

León Urrutia, M., Vázquez Cano, E., & López-Meneses, E. (2017). MOOC learning analytics using real-time dynamic metrics. *@tic. Revista D'innovació Educativa, 18*, 38-47. https://doi.org/10.7203/attic.18.10022

Lester, J., Klein, C., Johri, A., & Rangwala, H. (Eds.) (2018). *Learning analytics in higher education: Current innovations, future potential, and practical applications*. Routledge.

Liyanagunawardena, T. R., Adams, A. A., & Williams, S. A. (2013). MOOCs: A systematic study of the published literature 2008-2012. *International Review of Research in Open and Distributed Learning, 14*(3), 202-227. https://doi.org/10.19173/irrodl.v14i3.1455

Martín-Monje, E. (in press). Language MOOCs as an emerging field of research: From theory to practice. In M. M. Suárez & W. El-Henawy (Eds.), *Optimizing ONLINE English language learning and teaching* (pp. 25-43). London: Springer.

Martín-Monje, E., Castrillo, M. D., & Mañana-Rodríguez, J. (2018). Understanding online interaction in language MOOCs through learning analytics. *Computer Assisted Language Learning, 31*(3), 251-272. https://doi.org/10.1080/09588221.2017.1378237

Maseleno, A., Sabani, N., Huda, M., Ahmad, R., Jasmi, K. A., & Basiron, B. (2018). Demystifying learning analytics in personalised learning. *International Journal of Engineering & Technology, 7*(3), 1124-1129. https://doi.org/10.14419/ijet.v7i3.9789

Martín-Monje, E., & Barcena, E. (Eds.). (2014). *Language MOOCs: Providing learning, transcending boundaries*. De Gruyter Open.

Martín-Monje, E., & Borthwick, K. (2021). Researching massive open online courses for language teaching and learning. *ReCALL, 33*(2), 107-110. https://doi.org/10.1017/S0958344021000094

Martín-Monje, E., Castrillo, M. D., & Mañana-Rodríguez, J. (2018). Understanding online interaction in language MOOCs through learning analytics. *Computer Assisted Language Learning, 31*(3), 251-272. https://doi.org/10.1080/09588221.2017.1378237

Maya-Jariego, I., Holgado, D., González-Tinoco, E., Castaño-Muñoz, J., & Punie, Y. (2020). Typology of motivation and learning intentions of users in MOOCs: The MOOCKNOWLEDGE study. *Educational Technology Research and Development, 68*(1), 203-224. https://doi.org/10.1007/s11423-019-09682-3

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, *20*, 83-108. https://doi.org/10.1017/S0272263198001041

Mullaney, T. (2014). *Making sense of MOOCs: A reconceptualization of HarvardX courses and their students.* Retrieved from https://ssrn.com/abstract=2463736

Nurieva, G. R., & Garaeva, L, M. (2020). Zoom-based distance learning of English as a foreign language. *Journal of Research in Applied Linguistics, 11*, 439-448. https://doi.org/10.22055/rals.2020.16344

Panagiotidis, P. (2019). *MOOCs for language learning: Reality and prospects*. Society for Information Technology & Teacher Education International Conference.

del Peral Pérez, J. J. (2019). Patrones temporales de participación en MOOC. Estudio de un MOOC de lenguas. *RIED. Revista Iberoamericana de Educación a Distancia, 22*(2), 287-303. http://dx.doi.org/10.5944/ried.22.2.23109

del Peral Pérez, J. J. (2022). *El perfil del estudiante en los MOOC: identificación de patrones temporales de uso a partir del análisis de dos MOOC de lenguas.* [Unpublished doctoral dissertation, Universidad Nacional de Educación a Distancia (UNED)].

Poellhuber, B., Roy, N., & Bouchoucha, I. (2019). Understanding participant's behavior in massively open online courses. *International Review of Research in Open and Distributed Learning, 20*(1), 28-41. https://doi.org/10.19173/irrodl.v20i1.3709

Poy, Y., & Gonzales-Aguilar, A. (2014). Factores de éxito de los MOOC: algunas consideraciones críticas. *Revista Ibérica de Sistemas y Tecnologías de Información*, *SPE1*, 105-118. https://doi.org/10.4304/risti.e1.105-118

Read, T., & Sedano, B. (2021). The role of scaffolding in LMOOCs for displaced people. *Lengua y Migración/Language and Migration, 13*(2), 115-133.

Reich, J. (2014). MOOC completion and retention in the context of student intent. *EDUCAUSE Review*. Retrieved from http://www.educause.edu/ero/article/mooc-completion-and-retentioncontext-student-intent

Reich, J., & Ruipérez-Valiente, J. A. (2019). The MOOC pivot. *Science 363*, 130-13. https://doi.org/10.1126/science.aav7958

Romero, C., Cerezo, R., Bogarín, A., & Sánchez-Santillán, M. (2016). Educational process mining: A tutorial and case study using Moodle data sets. In S. ElAtia, D. Ipperciel, & O. R. Zaïane (Eds.), *Data mining and learning analytics* (pp. 1-28). London: Wiley.

Romero, C., & Ventura, S. (2020) Educational data mining and learning analytics: An updated survey. *WIREs Data Mining Knowl Discov, 10:e1355*. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1355

Sallam, M. H., Martín-Monje, E., & Li, Y. (2022). Research trends in language MOOC studies: A systematic review of the published literature (2012-2018). *Computer Assisted Language Learning, 35*(4), 764-791. https://doi.org/10.1080/09588221.2020.1744668

Sanz Gil, M. (2021). LMOOC para la integración de personas desplazadas. *Anales de Filología Francesa*, *29*, 463-477. https://doi.org/10.6018/analesff.483181

Sclater, N. (2017). *Learning analytics explained*. Routledge.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing, 97*, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Sunar, A., Abbasi, R., Davis, H., White, S., & Aljohani, N. (2020). Modelling MOOC learners' social behaviors. *Computers in Human Behavior, 107*, 50-62. https://doi.org/10.1016/j.chb.2018.12.013

Tatman, R. (2018). Data cleaning challenge: Scale and normalize data. *Kaggle*. Retrieved from https://www.kaggle.com/rtatman/data-cleaning-challenge-scale-and- normalize-data

Tavakoli, P., & Foster, P. (2011). Task design and second language performance: The effect of narrative type on learner output. *Language Learning, 61*, 37-72. https://doi.org/10.1111/j.1467-9922.2011.00642.x

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 6*3(2), 411-423. https://doi.org/10.1111/1467-9868.00293

Xu, B., & Yang, D. (2016). Motivation classification and grade prediction for MOOCs learners. *Computational Intelligence and Neuroscience, 4*, 1-7. https://doi.org/10.1155/2016/2174613

Zeng, S., Zhang, J., Gao, M., Xu, K. M., & Zhang, J. (2020). Using learning analytics to understand collective attention in language MOOCs. *Computer Assisted Language Learning, 35*(7), 1594-1619. https://doi.org/10.1080/09588221.2020.1825094

Zhang, Y., & Sun, R. (2023). LMOOC research 2014 to 2021: What have we done and where are we going next? *ReCALL*, 1-16. https://doi.org/10.1017/S0958344022000246