Profiling Language Learners in the Big Data Era

Mauro Ocaña
Universidad de las Fuerzas Armadas ESPE Ecuador
The University of Queensland
Australia

Hassan Khosravi
The University of Queensland
Australia

Aneesha Bakharia
The University of Queensland
Australia

The educational data revolution has empowered universities and educational institutes with rich data on their students, including information on their academic data (e.g., program completion, course enrolment, grades), learning activities (e.g., learning materials reviewed, discussion forum interactions, learning videos watched, projects conducted), learning process (i.e., time, place, path or pace of learning activities), learning experience (e.g., reflections, views, preferences) and assessment results. In this paper, we apply clustering to profile students from one of the largest Massive Open Online Courses (MOOCs) in the field of Second Language Learning. We first analyse the profiles, revealing the diversity among students taking the same course. We then, referring to the results of our analysis, discuss how profiling as a tool can be utilised to identify at-risk students, improve course design and delivery, provide targeted teaching practices, compare and contrast different offerings to evaluate interventions, develop policy, and improve self-regulation in students. The findings have implications for the fields of personalised learning and differentiated instruction.

Keywords: Big data, learning analytics, learner profiles, k-means clustering, online, language, IELTS.

1. Introduction

Big data are defined as "large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions" by the Oxford dictionary. As opposed to traditional data sets that are usually the result of long and intentional planning by the researcher, Big data are often automatically created by the interaction of users in every organisation at every size, and in every niche. This increase in the volume, velocity, variety and veracity referred to as the four Vs of Big data (Gantz & Reinsel, 2012) on user data has provided the opportunity for companies, governments, and individuals to record and analyse information pertaining to a user's individual, psychological and behavioural characteristics. This information can assist in constructing groups, referred to as profiles, of users who have similar characteristics. Profiling has been used in a wide range of domains such as medicine (Liu, 2018), banking (Schewe et al., 2002), marketing (Boe et al., 2001) and politics (Arian et al., 2017) to derive insight from large data sets.

With the recent advances in technology, education has grown from being a commodity of the few to being massified for the "transmission of skills" to being "universal" for a global population that needs to adapt to accelerated social and technological changes (Trow, 2007). Using video lectures at their core, Massive Open Online Courses (MOOCs) have emerged as an affordable solution in Higher Education to disseminate knowledge (Christensen et al., 2013). These days MOOCs have established themselves on the educational scene as a viable option for providing formal or informal training at scale. As the name implies, one of the defining characteristics of a MOOC is having a large number of students enrolled into the course from anywhere in the planet. With technologies reaching nearly "every corner of the world" (World Bank, 2018) enrolled students are very diverse across many demographic dimensions. A benefit of online education is that it captures students' data and their performed learning activities via e-learning systems, providing the ability to get detailed analytics and insights about the students and their learning process. In a recent trend, profiling methods have been applied to data collected via MOOCs (Ferguson & Clow, 2015; Khalil & Ebner, 2017; van den Beemt et al., 2018; Kovanović et al., 2017; Khosravi & Cooper, 2017). These works have been very promising providing insight on the diverse needs of the student population. Consequently, profiling students has been recognised as a desirable approach in the Big data era that can contribute to the facilitation of a more tailored learning experience for individual learners (Khalil & Ebner, 2017).

In this study, we first derive learning profiles for one of the largest language MOOCs existing to date (Cook, 2018), the IELTS Academic Test Preparation course developed by the University of Queensland and offered on the edX platform, which had approximately 272,187 learner enrolments in its first run between 2015-2016. The studied data set includes information about the students' demographics (e.g., age, gender, race), learning activities

(e.g., learning materials reviewed, discussion forum interactions, video-lectures watched, assessment items submitted) and learning process (i.e., time, place, path or pace of learning activities). As our work is more focused on the way students learn rather than their race, gender or age, these specific demographic traits have been omitted. We then, referring to the results of our analysis, discuss how profiling as a tool can provide meaningful benefits for different stakeholders involved in higher education. This will be especially helpful when trying to personalise or differentiate instruction.

2. Related Work

Profiling has a long history of being used in education even before the Big data era. For centuries, students have been profiled and consequently "educated in batches". In the 1970s, a range of competing and contested theories emerged that aim to profile learners based on their "learning styles" (Coffield et. al 2004, Kirschner, 2017). These theories invited teachers to use survey instruments to assess the learning style of their students and to adapt their teaching methods to best fit the needs of their students. Similarly, in language learning most, if not all, attempts to profile learning in the language field have heavily relied on surveys. One of the first to profile students was Stern (1975) who examined language learning strategies to profile the "good language learner". Later Oxford (1990, 1995) with her SILL (Strategy Inventory for Language Learning) profiled students based on the use of strategies. Another example includes a study by Muñoz and Singleton (2007) who created profiles of "*exceptional learners*" in speaking. Other studies have looked at profile differences between learners of different languages. For example, surveys show that users enrolled in less commonly-taught languages (e.g. Russian) have different profiles from those enrolled in commonly taught languages (e.g English). The former have, in general, previous knowledge of another language, study more for personal reasons rather than for complying with curricular demands and are older on average than the latter (Brown, 2009; Magnan, Murphy, Sahakyan, & Kim, 2012). In another less commonly taught language worldwide, Japanese, learners are asked about their instructional preferences to configure their own profile via survey so they can make a better use of the Strategy Inventory for Learning Kanji (SILK, found at http://kanji-silk.net).

With the emergence of data from MOOCs and large on-campus courses, development of student profiles has attracted the attention of researchers. In a highly cited study, Kizilcec, Piech, and Schneider (2013) found four profiles of engagement: *completing* (users completing most assessments items), *auditing* (learners who mostly watched video-lectures and did few assessment items) , *disengaging* (completed assessments only at the beginning of the course) and *sampling* (explored the content the first week). This study was later replicated by Ferguson and Clow (2016) bringing to attention the fact that despite rigour in methods, when analysing online behaviour some profiles can be similar across MOOCs and some cannot. In blended learning, Lust et al. (2013) used profiles to identify groups of *no-users, intensive users, selective users* and *limited users*. Brooks, Epp, Logan, & Greer (2011) found *minimal active learners, disillusioned learners, deferred learners* and *just-in-time learners*. Mirriahi, Liaqat, Dawson, & Gašević (2016) identified *minimalists, task focused, disenchanted* and *intensive learners*. Other studies show instructional preferences (instructor-led vs self-directed), attitude traits (Watson, Watson, Yu, Alamri, & Mueller, 2017) while Lynda (2017) used profiles to perform peer-assessment. In an engineering course, Khosravi and Cooper (2017) found sub-populations of students with extreme patterns of engagement: the *"overly engaged participants"* and the *"infrequent participants"*. Corrin, Barba, and Bakharia (2017) found five different learner profiles of students when help-seeking in MOOCs: *low engagement students, assessment-focused -low grades, passive engagement, active engagement, assessment-focused- high grades*. Reidsema et. al (2017) analysed the learning pathways of students in a large flipped engineering course and Kizilcec, Pérez-Sanagustín, and Maldonado (2017) profiled students who focused on specific strategies (help-seeking, goal-setting and strategic planning) of Self-regulated learning.

To the best of our knowledge, to date there are only two studies that have attempted to describe language learners on a large scale. Türkay (2017) used demographic information and self-reporting surveys of 100 online courses to discover motivational differences between English language learners (ELLs, learners who self-identify as *non-fluent* in English) and non-English language learners (non-ELLs, students who identify themselves as *fluent* in English). ELLs are "more motivated to earn a certificate" despite reporting a lack of interest in earning credit and are also said to be eager to engage with the online community despite their participation in forums being lower than that of non-ELLs. In a different study, Martín-Monje (2018), found that learners' favourite learning object in a MOOC was video-lectures and then, based on the combination of use of learning objects (article, video or book), that most learners were "*viewers*", who accessed content but did not submit tasks. In this paper, we focus on profiling students from one of the largest language learning MOOCs by taking a methodological approach that deals with multiple learning variables at the same time.

3. Research Methodology

In this section, the research methodology is presented. In Section 3.1 the IELTS MOOC is described. Section 3.2 describes the course assessment. In Section 3.3 student demographic data is explained. Section 3.4 describes the student event logs and grade data tracked by edX. Finally, Section 3.5 explains the profiling approach used to determine and analyse the different student profiles.

3.1 Course Overview

The IELTS Academic Test preparation course launched by The University of Queensland in edX in November 2015 is analysed in this paper. Each section of the course is divided into chapters, one for each language skill: Listening, Speaking, Reading and Writing. These chapters correspond to the sections of a real IELTS Academic test: Listening, Reading, Writing and Speaking. Each chapter then comprises video lectures that explain strategies to master the micro-skills assessed in the sections of a real IELTS test e.g (skimming, scanning, identifying paraphrases and references). Each chapter also includes practical exercises in various formats to put into practice the strategies explained.

3.2 Course Assessment

While the receptive macro skills (Listening and Reading) can be assessed objectively through the edX platform, the productive macro skills (Writing and Speaking) require each participant to compare their own performance against a set of rubrics. These factors have implications for the assessment tasks throughout the course which are reflected in the assignment policy that assigns 48% to Listening (24% for activities and 24% for the practice tests), 48% to Reading (24% for activities and 24% for practice test), 2% for Speaking self-assessment and 2% for Writing self- assessment.

3.3 Participants

A total of 272,187 users from 212 countries around the world enrolled in the course between November 2015 and November 2016. The overall median age of learners was 29, with most users falling into the age range 26-40 years old (60.7%) followed by a group aged under 25 (29.8%) and finally 41 and over (9.5%). The self-reported data also show that 50.8% held a Higher Education degree, 27.5% an Advanced degree (Doctorate, Master's or Professional degree) and 19.7% a High School diploma or less. To focus our analysis on students who made a serious attempt towards completion of the course, we limit our analysis to data from students who received a final grade of at least 20%. Therefore, the analysis includes data from 22,164 students.

3.4 Data Organisation

Data were obtained through the edX platform itself. Table 1 contains the list of features that was created for each student and provided to the k-means algorithm (see Section 3.5). The features have been grouped together as shown in Table 1. Some features represent aggregate counts (e.g. number of forum posts) while others require data pre-processing (e.g. average time between sessions and average number of chapters completed per session). The features have been selected to encode visitor frequency (average number of sessions per week), time spent on task (average session duration), how learners viewed and reviewed video (number of plays, number of pauses), and how learners completed course content (average number of chapters completed per session). The average number of sessions spent on each chapter is included to give an indication of how learners were distributing their time on these four skills.

Table 1: Features created for each IELTS course learner

Feature types	Descriptions
Sessions	s_1 = average session duration time, s_2 = total number of sessions, s_3=time between sessions.
Video interactivity	v_1 = number of plays, v_2 = number of pauses, v_3 = number of video seeks, v_4 = number of times a transcript was viewed.
Community engagement	e_1 = number of forum posts read, e_2 = number of comments posts, e_3 = number of forum votes
Content	c_1 = number of sessions which include access to chapter 1 (Listening), c_2 =number of sessions which include access to chapter 2 (Speaking), c_3 =number of sessions which include access to chapter 3 (Reading), c_4 =number of sessions which include access to chapter 4 (Writing)
Assessment	a_1 = number of problems attempted, a_2 = first summative assessment, a_3=second summative assessment
Final Grade	g_1 = 1st quartile, g_2 = median, g_3 = 3rd quartile

3.5 Profiling Approach

As per previous studies (e.g. Khosravi & Cooper, 2017), k-means clustering was used to find student Learning Profiles. K-means clustering is an unsupervised algorithm capable of finding groups of students with similar characteristics. It takes as input a matrix, each row representing an individual, and aggregates associated features as columns in the matrix. The selection of appropriate features is very important and is known as feature engineering. The features included in this study have been specifically designed to reveal learner similarity from a personalised learning perspective. The k-means algorithm requires that the number of clusters (i.e., student profiles) be provided as a parameter. The clustering algorithm was run 100 times to select the solution with the highest likelihood. To determine an appropriate value for the number of clusters in the data set the elbow method was used. The elbow method computes the sum of within-cluster variances which can then be plotted in a curve. The most prominent turning point in the curve suggests the best number of clusters. Within this paper each cluster is referred to as a student profile and analysed.

4. Data analysis

This section analyses the learner population which took the IELTS Academic Test preparation course launched by The University of Queensland in edX in November 2015 by applying the methodology presented in Section 3. The results obtained from running k-means reported five clusters also known as profiles. These clusters are ordered from C1 to C5 in descending population size as shown in Table 2.

4.1 Cluster-based analysis

A short description of the resulting clusters is provided below. All of the reported numbers refer to average values for the entire cluster and not any individual.

Strong starters, weak finishers (C1): The largest cluster, containing 38.86% of the analysed population, gave more emphasis to the first section presented in the course (Listening), visiting it more than other sections and getting high scores only in the corresponding formative assessment, then exhibiting a gradual decrease in participation and a sharp drop in grades. They did very well in the formative assessment of the first section where strategies (e.g identifying paraphrases or predicting words and situations) in a listening context were provided. In turn, they performed relatively well in the corresponding summative assessment. In comparison with other chapters, they had a higher level of engagement with the content of the first chapter. These learners rarely engaged with the online community and had a very low average number of forum reads (2.49) and even a lower average number of forum posts (0.06) - meaning that many members never posted.

More content, less assessment (C2): The second largest cluster, containing 19.36% of the analyzed population, had the particularity of engaging well with the content spread throughout the course by visiting each of the four sections (skills) uniformly and making high use of the video features (e.g pauses, seeks, speed changes, show

transcripts). Despite this, they did not seem very interested in the practice tests that represent the summative assessment for each section, but only in the formative tasks for the two receptive skills (Listening and Reading) which were assessed objectively (in multiple choice format). Both their average number of sessions and their average session length were high throughout the course. They were also prompt to return between one session and the next (290336 s). Their forum reads are less than moderate (4.45) with a very low number of forum posts (0.12).

More assessment, less content (C3): The third largest cluster, containing 17.87% of the analyzed population, has the lowest level of engagement; members of this cluster had the lowest average number of sessions (9.08) with the lowest average session length (1258.33 s) and the lowest average number of video plays. Their community engagement was also lowest of all clusters both in their forum reads (1.25) and their forum posts (0.03). Interestingly, they scored high in the first summative assessment presented in the course (Listening). This might indicate that this group had minimal interest in the content of the course and in their short time spent on the course mostly focused on the assessment.

Very high engagement, moderate performance (C4): The fourth largest cluster, containing 16.03% of the analyzed population, has the highest number of sessions (41.30) with the highest rates of video interactivity (e.g video seeks, video speed changes, show transcripts) of all the groups. They also interacted steadily with the content in each section of the course. Having the highest average session length (2561.59 s) and the lowest average time between sessions in comparison with other clusters, learners in this cluster were quicker to come back to the course than the other clusters. They had the second highest average grade; they performed highly in the formative (and objective) assessment of the receptive skills (Listening and Reading), moderately well on the summative assessments of both Speaking and Writing which are subjective (open answers format) and constitute only 4% of assessment overall (2% for each productive skill). Compared to other clusters, their participation in forums was neither high nor low: reads (4.45) and forum posts (0.14).

High engagement, high performance (C5): The smallest cluster, containing only 7.87% of the analyzed population, belongs to those learners who got the highest scores of all. They outperformed the other clusters in nearly all the features performing very well across the formative and summative assessment throughout the four skills and exhibited other positive characteristics in assessment-related events such as check progress, show answers and attempt problems. They had the highest number of play and pause videos counts as well as other video features (e.g., seeks, stops, show transcripts), indicating that they were more actively involved learners while watching the videos. They displayed a very high number of sessions (39.07) with the highest average session length (2604.16) though their average returning time between sessions is not the highest among other clusters. They have the highest number of forum reads (13.06) and forum posts (0.48) among all of the clusters.

Table 2: Using k-means to cluster the class population across features described in Table 1

label	Size	Sessions		Video Interactivity				Community engagement			Content				Assessment engagement			Grade		
		s_1	s_2	v_1	v_2	v_3	v_4	c_1	c_2	c_3	c_1	c_2	c_3	c_4	a_1	a_2	a_3	g_1	g_2	g_3
C1	38.8%	2128	18	104	57	24	7	2	0.02	0.01	20	7.5	2.6	1	372	0.57	0.05	0.29	0.38	0.44
C2	19.3%	2566	35	304	152	80	17	4	0.06	0.02	20	16	22	10	634	0.31	0.08	0.31	0.41	0.5
C3	17.8%	1258	9	41	19	9	2	1	0.01	0.01	6	3.1	2.3	1	84	0.75	0.22	0.21	0.24	0.31
C4	16.%	2561	41	331	168	89	22	6	0.06	0.04	22	18.9	24	11	789	0.77	0.64	0.67	0.77	0.86
C5	7.8%	2604	39	410	176	82	20	13	0.23	0.31	21	19.5	20.7	12	921	0.81	0.76	0.80	0.89	0.95

Figure 1 visually illustrates how the five clusters compare against one another across some of the main features obtained through the edX platform.

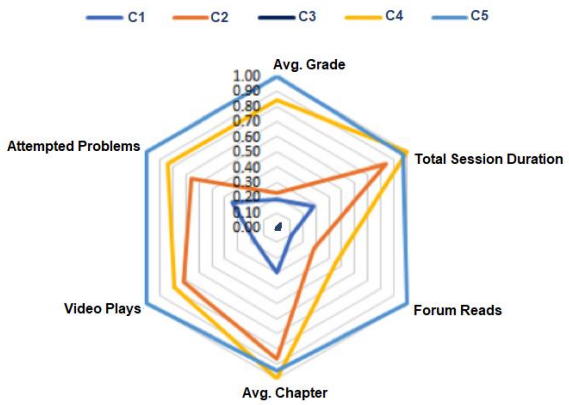

Figure 1: A visual illustration of how different clusters compare against some of the main features that were introduced in Table 1

5. Benefits of Profiling Students

In this section, we discuss the potential benefits of profiling for different stakeholders. The benefits that arise from profiling students are mainly due to the affordances provided by the clustering algorithm (i.e. k-means). The important properties of the k-means clustering algorithm include the ability to find groups of similar students even when a large number of features are provided to the algorithm and the ability to assign each student to a profile. These two properties allow statistical summaries to be calculated for each cluster, which aids in the interpretation and naming of profiles. Some of the main benefits of profiling are discussed below.

Identifying at-risk students: Methods of identifying at-risk students with the aim of utilising retention strategies have been well studied in the literature Marbouti et al. (2016). Profiling students at early stages in the semester to identify disengaged students can be used as a viable option for identifying at-risk students (De Paepe et al., 2016). In our study, students assigned to C1 may be considered as at-risk students.

Improving course design and delivery. The profiles provide detailed information regarding the engagement and performance of students throughout the course, which may be used towards improving the design and delivery of a course. For example, a high number of pauses or seeks on some videos across one more clusters may suggest that students find the content of the video challenging or confusing. This information may be used towards re-evaluating the quality and consequently updating that video. Profiles can also provide insightful information in terms of course delivery. For example, in our study, the students associated with the *"More assessment, less content"* profile seem to aim to attempt assessment items without first going through the associated learning material. Once this phenomenon is identified, it is possible to change the course delivery mechanisms to minimise this behaviour. For example, the assessment items can be embedded in the learning material to encourage students to review the learning content before attempting the assessment items.

Provide targeted student interventions. Profiles may be used to provide targeted interventions for students associated with each cluster based on their behaviour or learning needs. For example, an instructor may wish to share optional additional advance learning material with students in the "*High engagement, high performance*" cluster while providing more support material and words of encouragement for students in the "*Strong starters, weak finisher*" cluster.

Comparing offerings and evaluate interventions. Profiles can be used to visually compare and contrast different courses or different course offerings. For example, it is possible to visually compare profiles of two offerings of the same course to determine how the clusters are similar or different in terms of students' engagement and performance. This may be used as a mechanism to evaluate interventions. For example, if the two offerings are using a different set of learning material (e.g. videos), it is possible to evaluate and visually determine which set of videos have led to better engagement and performance.

Developing policy. Based on reports of learning profiles from across an institute, university administrators may have a global view of the effectiveness of an action or an intervention, which may lead to the development of policies. For example, in the 2015-2016 offering of this IELTS course, access to assessment items was available to both paid and non-paid users. In the 2016-2017 offering of this IELTS course, access to assessment items was

only available to paid users. Comparison of the profiles across many MOOCs that have tried out features to be included or excluded for non-paying users may enable university administrators to develop policy around access.

Promoting self-regulation. Sharing the profiles with students enables them to be aware of their strengths and weaknesses so that they themselves can suggest the best mechanisms to overcome their flaws, decide which paths to take and even become knowledgeable enough to create their own cognitive tools.

Table 3 shows how diverse stakeholders within an educational ecosystem are able to use student profiles for a range of tasks.

	University Administrators	Program Administrat.	Instructors	Learning Designers	Educational Researchers	Students
Identify at-risk students	**X**	**X**	**X**		**X**	**X**
Improve course design and delivery			**X**	**X**	**X**	
Provide targeted student interventions, scaffolded instruction and feedback			**X**	**X**	**X**	**X**
Compare offerings and evaluate interventions		**X**	**X**	**X**	**X**	
Develop policy	**X**	**X**			**X**	
Promote self-regulation			**X**		**X**	**X**

Discussion and Conclusion

This paper presents learning profiles of language test-takers as a means to identify who they were, not in terms of traditional profiling features such as age or country of origin (that may be misleading when assisting a learner) but in terms of actual behaviours when learning. Of particular interest are those behaviours which reflect weaknesses or needs during the learning process. They should be interpreted as a call to action for educational stakeholders to intervene.

Our results, reiterating findings from past studies (e.g. Ferguson & Clow, 2015; Khalil & Ebner, 2017; Khosravi & Cooper, 2017), suggest that learners are very diverse in terms of their approach, behaviour and performance. 38% of the analysed population were profiled as "*strong starters, weak finishers*" due to their high engagement at the beginning of the course and low engagement towards the end of the course. 19% of the analysed population were placed in the "*More content, less assessment*" profile as they primarily focused on watching videos and reviewing notes without engaging with the assignments. In contrast, 18% of the analysed population were placed in the "*More assessment, less content*" profile as they show no interest in the content and moved straight to the tests. 16% of the analysed population were profiled as having "*very high engagement, moderate performance"* and finally 8% of the analysed is profiled as having "*High engagement, high performance*".

In general, it can be said that the higher the engagement, the higher the grade. For example, clusters C5 and C4, which achieved the highest grades, also recorded the highest figures relating to features such as number of sessions, number of chapters covered, video plays and attempted problems. Of particular importance is cluster C2 which, despite having good engagement with the whole course, did not seem to be especially interested in assessment. In contrast, cluster C3 showed minimal interest in the content and focused their efforts mostly on the practice test. Learners in C3, were mainly using the MOOC to practise their IELTS skills and prepare for the official IELTS test with little motivation in obtaining a certificate from edX.

While some of the student clusters share some traits with others from past studies (e.g those highly engaged learners) due to the nature of the course there are also distinctive learner characteristics that stand out in this study. Student characteristics exhibited in each learning profile were the result of learning behaviours revealed throughout the course. This way of profiling students makes it a suitable fit to advance the field of personalised

education. Technology designers, educators and administrators all together may harness data captured by learning profiles to improve mechanisms that support those learners who fall behind, keep encouraging those who are doing well and keep all the others in between on track. Given the diversity among learners, we discussed how profiling as a tool can provide benefits for university administrators, program administrators, instructors learning designers, educational researchers and students. These benefits include identifying at-risk students, improving course design and delivery, providing targeted teaching practices, comparing and contrasting different offerings to evaluate interventions, developing policy, and improving self-regulation in students.

References

Arian, A., Krouwel, A., Pol, M., & Ventura, R. (2017). The election compass: party profiling and voter attitudes. In The Elections in Israel 2009 (pp. 275-298). Routledge. https://doi.org/10.4324/9781351297608-15

Boe, B. J., Hamrick, J. M., & Aarant, M. L. (2001). U.S. Patent No. 6,236,975. Washington, DC: U.S. Patent and Trademark Office.

Brown, A. (2009). Less Commonly Taught Language and Commonly Taught Language students: A demographic and academic comparison. Foreign Language Annals, 42(3), 405-423 https://doi.org/10.1111/j.1944-9720.2009.01036.x

Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D., & Emanuel, E. (2013). The MOOC phenomenon: who takes massive open online courses and why?. Available at SSRN 2350964. https://doi.org/10.2139/ssrn.2350964

Brooks, C., Epp, C. D., Logan, G., & Greer, J. (2011). The who, what, when, and why of lecture capture. Proceedings of the First International Conference on Learning Analytics and Knowledge, 86-92. https://doi.org/10.1145/2090116.2090128

Coffield, F., Moseley, D., Hall, E., & Ecclestone, K. (2004). Learning styles and pedagogy in post-16 learning: A systematic and critical review. In: Learning and Skills Research Centre London.

Corrin, L., Barba, P. G. d., & Bakharia, A. (2017). Using learning analytics to explore help-seeking learner profiles in MOOCs. Proceedings of the Seventh International Learning Analytics Knowledge Conference, 424-428. https://doi.org/10.1145/3027385.3027448

Cook, M. (2018). The 50 Most Popular MOOCs of All Time. Retrieved from https://www.onlinecoursereport.com/the- 50-most-popular-moocs-of-all-time/

De Paepe, L., Zhu, C., & DePryck, K. (2016). Drop-out, Retention, Satisfaction and Attainment of Online Learners of Dutch in Adult education. Paper presented at the E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2016, Washington, DC, United States.

Ferguson, R., & Clow, D. (2015). Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs). In Proceedings of the fifth international conference on learning analytics and knowledge (pp. 51-58). https://doi.org/10.1145/2723576.2723606

Ferguson, R., & Clow, D. (2016). Consistent commitment: Patterns of engagement across time in Massive Open Online Courses (MOOCs). Journal of Learning Analytics, 2(3), 55-80. https://doi.org/10.18608/jla.2015.23.5

Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, 2007(2012), 1-16.

Khalil, M & Ebner, M 2017, 'Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories', Journal of Computing in Higher Education, vol. 29, no. 1, pp. 114-132. https://doi.org/10.1007/s12528-016-9126-9

Khosravi, H., & Cooper, K. M. L. (2017). Using learning analytics to investigate patterns of performance and engagement in large classes. Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, 309-314. https://doi.org/10.1145/3017680.3017711

Kirschner, P. A. (2017). Stop propagating the learning styles myth. Computers & Education, 106, 166-171. https://doi.org/10.1016/j.compedu.2016.12.006

Kizilcec, R., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. Proceedings of the Third International Conference on Learning Analytics and Knowledge, 170-179. https://doi.org/10.1145/2460296.2460330

Kizilcec, R., Pérez-Sanagustín, M., & Maldonado, J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. Computers and Education, 104, 18-33. https://doi.org/10.1016/j.compedu.2016.10.001

Kovanović, V., Joksimović, S., Poquet, O., Hennis, T., Dawson, S., Gašević, D., & Siemens, G. (2017, March). Understanding the relationship between technology use and cognitive presence in MOOCs. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference (pp. 582-583). ACM. https://doi.org/10.1145/3027385.3029471

Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G. E., Smith, J. L., Mohtashamian, A., Stumpe, M. C. (2018). Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection. Archives of Pathology & Laboratory Medicine. https://doi.org/10.5858/arpa.2018-0147-OA

Lynda, H., Farida, B. D., Tassadit, B., & Samia, L. (2017). Peer assessment in MOOCs based on learners' profiles clustering. 2017 Eighth International Conference on Information Technology, 532-536. https://doi.org/10.1109/ICITECH.2017.8080054

Lust, G., Elen, J., & Clarebout, G. (2013). Regulation of tool-use within a blended course: Student differences and performance effects. Computers & Education, 60(1), 385-395. https://doi.org/10.1016/j.compedu.2012.09.001

Magnan S., Murphy, D., Sahakyan, N., & Kim S. (2012). Student Goals, Expectations, and the Standards for Foreign Language Learning. Foreign Language Annals, 45(2), 170-192. https://doi.org/10.1111/j.1944-9720.2012.01192.x

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education, 103, 1-15. https://doi.org/10.1016/j.compedu.2016.09.005

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

Mirriahi, N., Liaqat, D., Dawson, S., & Gašević, D. (2016). Uncovering student learning profiles with a video annotation tool: reflective learning with and without instructional norms. Educational Technology Research and Development, 64(6), 1083-1106. https://doi.org/10.1007/s11423-016-9449-2

Muñoz, C., & Singleton, D. (2007). Foreign accent in advanced learners: Two successful profiles. EUROSLA Yearbook, 7(1), 171-190. https://doi.org/10.1075/eurosla.7.10mun

Oxford, R. (1990). Language learning strategies: what every teacher should know: Newbury House Publisher.

Oxford, R. L., & Burry-Stock, J. A. (1995). Assessing the use of language learning strategies worldwide with the ESL/EFL version of the Strategy Inventory for Language Learning (SILL). System, 23(1), 1-23. https://doi.org/10.1016/0346-251X(94)00047-A

Reidsema, C., Khosravi, H., Fleming, M., Kavanagh, L., Achilles, N., & Fink, E. (2017) Analysing the learning pathways of students in a large flipped engineering course.

Schewe, K. D., Kaschek, R., Matthews, C., & Wallace, C. (2002, October). Modelling web-based banking systems: Story boarding and user profiling. In International Conference on Conceptual Modeling (pp. 427-439). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-45275-1_37

Stern, H. H. (1975). What can we learn from the good language learner? Canadian Modern Language Review, 31(4), 304-318. https://doi.org/10.3138/cmlr.31.4.304

Trow, M. (2007). Reflections on the transition from elite to mass to universal access: Forms and phases of higher education in modern societies since WWII. In J. J. F. Forest & P. G. Altbach (Eds.), International handbook of higher education (pp. 243-280). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-4012-2_13

Türkay, S., Eidelman, H., Rosen, Y., Seaton, D., Lopez, G., & Whitehill, J. (2017). Getting to Know English Language Learners in MOOCs: Their Motivations, Behaviors, and Outcomes. Paper presented at the Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale. Cambridge, Massachusetts, USA https://doi.org/10.1145/3051457.3053987

van den Beemt, A., Buijs, J., & van der Aalst, W. (2018). Analysing structured learning behaviour in massive open online courses (MOOCs): an approach based on process mining and clustering. International Review of Research in Open and Distributed Learning, 19(5). https://doi.org/10.19173/irrodl.v19i5.3748

Watson, S. L., Watson, W. R., Yu, J. H., Alamri, H., & Mueller, C. (2017). Learner profiles of attitudinal learningin a MOOC: An explanatory sequential mixed methods study. Computers & Education, 114, 274-285. https://doi.org/10.1016/j.compedu.2017.07.005

World Bank. (2018). Technology is changing the world of work: What does that mean for learning? In World Development Report 2018: Learning to Realize Education's Promise. Retrieved from https://openknowledge.worldbank.org/bitstream/handle/10986/28340/9781464810961_Spot05.pdf

Please cite as: Ocana, M., Khosravi, H. & Bakharia, M. (2019 Profiling Language Learners in the Big Data Era. In Y. W. Chew, K. M. Chan, and A. Alphonso (Eds.), *Personalised Learning. Diverse Goals. One Heart. ASCILITE 2019 Singapore* (pp. 237-245).