Journal of Computing in Higher Education (2020) 32:162–181
https://doi.org/10.1007/s12528-019-09231-0

Investigating variation in learning processes in a FutureLearn MOOC

Saman Rizvi[1] · Bart Rienties[1] · Jekaterina Rogaten[2] · René F. Kizilcec[3]

Published online: 14 June 2019

Abstract

Studies on engagement and learning design in Massive Open Online Courses (MOOCs) have laid the groundwork for understanding how people learn in this relatively new type of informal learning environment. To advance our understanding of how people learn in MOOCs, we investigate the intersection between learning design and the temporal process of engagement in the course. This study investigates the detailed processes of engagement using educational process mining in a FutureLearn science course (N = 2086 learners) and applying an established taxonomy of learning design to classify learning activities. The analyses were performed on three groups of learners categorised based upon their clicking behaviour. The process-mining results show at least one dominant pathway in each of the three groups, though multiple popular additional pathways were identified within each group. All three groups remained interested and engaged in the various learning and assessment activities. The findings from this study suggest that in the analysis of voluminous MOOC data there is value in first clustering learners and then investigating detailed progressions within each cluster that take the order and type of learning activities into account. The approach is promising because it provides insight into variation in behavioural sequences based on learners' intentions for earning a course certificate. These insights can inform the targeting of analytics-based interventions to support learners and inform MOOC designers about adapting learning activities to different groups of learners based on their goals.

Keywords MOOCs · Educational process mining · Learning design

✉ Saman Rizvi
saman.rizvi@open.ac.uk

Extended author information available on the last page of the article

Introduction

Open-access learning environments such as Massive Open Online Courses (MOOCs) attract people with a wide range of interests and learning objectives, which is reflected in the degree and nature of engagement with the learning content (Milligan and Littlejohn 2017; Kizilcec and Schneider 2015). However, participation levels and assessment outcomes alone do not constitute robust evidence of learning or academic success writ large (Henderikx et al. 2017; Joksimović et al. 2017). While early research on MOOCs focused on understanding completion rates and final course grades, more recent work has examined how learners are moving through the course content as a way of understanding the learning process itself.

Regardless of whether a learner completes a MOOC, academic success or failure may be partly hidden in their journey through the learning activities in the course (Rizvi et al. 2018, 2019). Given the processual nature of learning, we can investigate learning by measuring detailed interactions with learning activities, such as videos, assessments, and interpersonal exchanges, and analysing learners' progression through these activities (Davis et al. 2016; Maldonado-Mahauad et al. 2018). Unlike face-to-face or blended learning environments, online courses are instrumented such that learner interactions are recorded in voluminous system logs, offering an unprecedented granularity for studying learning at scale. Educational research on log-based behavioural modelling in Intelligent Tutoring System (ITS) and Learning Management Systems (LMS) has found that log-based analyses can provide deep insights into how learners engage and interact with different learning activities (Bogarín et al. 2018; Sonnenberg and Bannert 2015). Yet despite increasing efforts to advance learning science research with log-based analyses in formal and blended learning environments, more research is needed to advance our understanding of learning processes in online learning environments (Bogarín et al. 2018; Juhaňák et al. 2017).

To advance an understanding of learner behaviour in MOOCs, studies have used clustering techniques to identify learner subpopulations based upon their overall resource-engagement behaviour (Li and Baker 2018; Ferguson and Clow 2015; Kizilcec et al. 2013), and more recently sequence-mining techniques to identify common engagement sequences that may reflect learning processes (Davis et al. 2018; Guo and Reinecke 2014). In order to understand learning processes in MOOCs, findings from these studies suggest that it helps to first group learners based on their general behavioural profile to reduce variance due to different enrolment intentions, and then to examine fine-grained interaction processes with the learning activities.

While these sequence-mining techniques have provided important insights in how different groups of learners engage in MOOCs, some researchers have argued that these approaches need to be embedded in strong learning science principles (Mangaroska and Giannakos 2018; Winne 2017). Indeed, the design of the online learning environment is known to influence learners' progression in different types of learning activities (Nguyen et al. 2018; Rienties and Toetenel

2016). Success in online learning has been found to be closely linked to learning design, which is defined as the process of designing pedagogically informed learning activities to support learners while remaining aligned with the curriculum (Conole 2012). Yet research on the pedagogical learning design of MOOCs is at an early stage (Davis et al. 2016; Sergis et al. 2017). We adopt learning design as a lens for investigating learners' interaction processes with the goal of finding empirical support for actionable recommendations to course designers and policymakers who have control over the learning design.

The present research reports on our implementation and evaluation of this approach by combining both sequence-mining techniques with learning design approaches to better understand how and why groups of learners engage in a science MOOC over time. In particular, our current implementation extends prior work that has identified three primary clusters of engagement in courses offered on the FutureLearn platform (Rizvi et al. 2018). The clustering was based on the degree to which learners marked activities in the course as completed: "Markers" are learners who marked all their activities as completed; "Partial-Markers" are those who marked only a few activities, and "Non-Markers" marked none of their activities as completed. For each of these groups, we investigate detailed processes of engagement with the learning activities according to an established taxonomy of course activities in the learning design. The findings of this study can inform approaches to adapting course content and learning activities in particular to different groups of learners based on their learning goals.

Literature review

The intrinsic features of MOOCs make them accessible to diverse populations of learners. This allows for a spectrum of learning approaches and contexts, including a variety of languages, cultural settings, pedagogical strategies, and technologies (Jansen and Schuwer 2015; Morgado et al. 2014). In comparison to other online learning environments, MOOC learning environments are not only "open" but often require learners to be highly self-directed and self-regulated (Maldonado-Mahauad et al. 2018). For MOOC design and development, a variety in content types have been recommended, moving away from the predominantly video-based courses (Jansen and Schuwer 2015). The essential features of MOOCs facilitate learners with a mediated experience: i.e., fewer constraints for time, distance, prerequisites or technological barriers (Sparke 2017; Kizilcec et al. 2017). This "structured-informality" makes MOOCs unique, and different from formal residential learning, even from traditional distance or online learning, and opens doors to large-scale adoption. Our current study is an attempt to understand how the learning design of MOOCs might impact the way learners engage and progress in the course.

Learning design

In his seminal work, Mayer (2005) wrote that learning comprised of the active processes of filtering, selecting, organising, and integrating new information. At present, MOOCs developers like FutureLearn, Coursera, and edX seem to optimise the design of MOOCs to increase study success (i.e. completion rates), and to lessen the so-called cognitive load for learners by adjusting topic difficulty and information or task presentation, the robustness of acquired knowledge. By making the acquisition of textual, visual or auditory information natural and easy for learners, MOOC providers aim not only to attract but also retain more learners (Sergis et al. 2017; Rai and Chunrao 2016; Margaryan et al. 2015). Additionally, it is common that learners distribute their time to different learning activities to get the maximum (subjective) benefit within a limited time frame (Maldonado-Mahauad et al. 2018; Wigfield and Eccles 2000). Therefore, the structural constructs (i.e., learning activities) of MOOCs need to be in alignment with respective learning objectives. Thus, the temporal dynamics of designed learning activities are of special interest to researchers and MOOC developers.

Learning design (LD) can be defined as the process of designing pedagogically informed learning activities to support learners while remaining aligned with the curriculum. In a MOOC, LD can provide a consistent way to map individual learning activities. This study has theoretical groundings in the conceptual framework for Learning Design recommended by The OU Learning Design Initiative (OULDI) project (Cross et al. 2012). This conceptual framework provides a foundation for the MOOC designs at FutureLearn platform (Sharples 2015), which is the primary source of MOOC data in this research.

The formal taxonomy for OULDI, shown in Table 1, was developed by Conole (2012). LD has been described as reusable, adaptable description or template which aims to "make the structures of intended teaching and learning—the pedagogy—more visible and explicit thereby promoting understanding and reflection" (Cross and Conole 2009). Reusability, adaptability, and abstraction of the overall course structure are few of the strengths of OULDI. This proposed taxonomy provides a way to abstract different learning activities in a meaningful way. It suggests that all learning tasks can be categorised as one of seven activity types.

In formal online learning contexts the impact of LDs on learners' behaviour, satisfaction, and learning outcomes has been widely acknowledged (Rienties and Toetenel 2016). Likewise, Nguyen et al. (2017) found preliminary support of the impact of LD on learners' online engagement, whereby "LD could explain up to 60% of the variance of the time spent on VLE platform". However, most of the research on LD and learning focused on measures of learning that are not processual (Mangaroska and Giannakos 2018). For example, the impact of LDs on learning outcomes or overall engagement has been analysed by a study of Rienties and Toetenel (2016), but without taking consideration of processual nature of learning. In other words, the OULDI framework has been empirically tested in large-scale studies (Nguyen et al. 2017; Nguyen 2017), but not in informal

Table 1 Learning design taxonomy

	Type of activity	Example
Assimilative	Attending to information	Read, watch, listen, think about, access
Finding and handling information	Searching for and processing information	List, analyse, collate, plot, find, discover, access, use, gather
Communication	Discussing module related content with at least one other person (student or tutor)	Communicate, debate, discuss, argue, share, report, collaborate, present, describe
Productive	Actively constructing an artefact	Create, build, make, design, construct, contribute, complete
Experiential	Applying learning in a real-world setting	Practice, apply, mimic, experience, explore, investigate
Interactive/adaptive	Applying learning in a simulated setting	Explore, experiment, trial, improve, model, simulate
Assessment	All forms of assessment (summative, formative and self-assessment)	Write, present, report, demonstrate, critique

learning settings and FutureLearn MOOCs in particular. In the current study, we employ OULDI to investigate the cognitive and pedagogical features of a FutureLearn MOOC in relation to learners' engagement and learning progression.

MOOC event logs and learning processes

Learning in MOOC environments produces large volumes of data, irrespective of how a MOOC has been designed. These data are produced from multiple sources, in a variety of formats, and with different levels of granularity (Romero and Ventura 2013). Within MOOCs, "trace data" or "clickstream data" are typically captured at a very fine-grained level. This participation log data presumably can be considered as a set of silent, passive observations. The volume of data increases immensely as we move from general course-related details to learner-related information. The data size increases even more if we go deeper into each learner's progress, from their learning sessions to individual learning activities accessed within those sessions (Fig. 1).

Stored log data have no inherent meaning per se, as clicking data does not necessarily mean behavioural engagement, let along cognitive processing or learning (Winne 2017). Indeed, Selwyn (2015) argued that the focus on these clicking data could lead to "dataveillance", and perhaps more importantly to a reductionist nature of data-based representation of diverse learners. Nonetheless, a substantial body of literature is emerging that suggests these clicking data streams, if used sensitively and sensibly, could provide important insights into how some groups of learners are engaging in MOOCs, while others might not be. Still, to date, only a small fraction of that data have been explored in extensive, systematic MOOC research (Bogarín et al. 2018; Winne 2017; Joksimović et al. 2017). In other words, there is still a paucity in systematic research exploring what aspects of these data are relevant and helpful in understanding learning processes (Winne 2017; Sparke 2017).

Learning can be assessed in a variety of ways, ranging from the learning outcomes like grades and certifications (Baker et al. 2016; Wang et al. 2015; Wen and Rosé 2014), to conceptualising learning as a process (Bogarín et al. 2018; Maldonado-Mahauad et al. 2018). While assuming learning as a process, several studies have recently explored log data to understand learners' progress, or processual

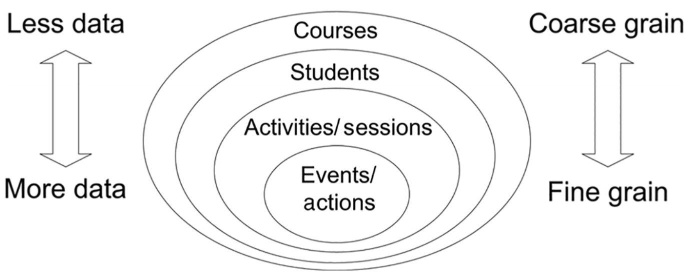

Fig. 1 Different levels of granularity and their relationship to the amount of data. *Source*: Romero and Ventura (2013)

learning, in different MOOC activities (Davis et al. 2016; Guo and Reinecke 2014; Kizilcec et al. 2013). For instance, to understand learners progression in Coursera MOOCs, (Kizilcec et al. 2013) used engagement patterns to categorise learners into four categories: completing (completed majority of the assessments), auditing (watched most of the videos but completed assessments infrequently), disengaging (completed assessments at the start of the MOOC, then gradually disengaged). Ferguson and Clow (2015) replicated this method in the context of FutureLearn MOOCs, whereby FutureLearn allows learners to specifically mark activities as 'complete'. Ferguson and Clow (2015) suggested that marking few or all activities as 'completed' signified a certain level of activity-engagement or learning commitment. Also, such clicking behaviour indicated a strategic way of getting a certificate.

Similarly, in a large-scale study of four edX MOOCs (Guo and Reinecke 2014, p. 6) found that participants exhibited a pattern of 'non-linear navigation through the course materials'. In particular, it was reported that so-called "certificate-earners" remained inclined towards the application of non-linear navigation strategies, whereby "certificate earners repeated visiting prior sequences three times as often, presumably to review older content." (Guo and Reinecke 2014, p. 6). Hence, this research suggested distinct navigational strategies, and that clicking (or not clicking) activities as "completed" represented two distinct psychological dispositions: one when a learner might be inclined to attain a certificate; and the other when learner showed no intention to get a certificate, yet, continued to learn.

Along the same lines, several authors (Davis et al. 2016; Guo and Reinecke 2014; Wen and Rosé 2014) have inspected MOOC learning sequences (or learning processes) in connection to assessment results, inclination towards certification, learning strategies or habits. For example, Wen and Rosé (2014) quarried transitions between two activities and linked the findings with behavioural patterns. A relatively similar approach of using two-step transition to map navigational strategies was used in the work of Guo and Reinecke (2014). Both studies found that generally learners progressed linearly, but certificate earners were more inclined to follow unstructured paths. Recently, a slightly different method was used by Davis et al. (2016), who studied MOOC learners' motivations, like binge (video) watching or 'quiz checking' (i.e., checking the quiz answers without attempting the quiz first). To capture the complexities of such motivations, the authors used eight-step long subsets of overall learning sequences. Their findings suggested that learners' progression through activities and the frequency with which they accessed various learning activities should be seen in the context of their inclination towards certification.

Given that our study is situated in the FutureLearn environment, it is noteworthy that FutureLearn's policy on "certificate of participation" allows for a non-linear navigation through the activities. In most courses, a learner must mark at least 50% of the course steps as complete and attempt every test question to get a certificate of participation. An initial analysis (Rizvi et al. 2018) of log data used in current study pointed towards three distinct clicking patterns, potentially representing three unique dispositions: Markers (i.e., those who marked all their activities as completed); Partial-Markers (i.e., those who marked few of the activities they assessed), and Non-Marker (i.e., those who never marked any of their activities as completed). This learners' grouping is unique and so is the MOOC designs offered via FutureLearn platform. Nonetheless, this categorisation is informed by similar categorisation stated in previous MOOC literature (Kizilcec et al. 2013; Ferguson and Clow 2015).

Apart from understanding the similar or dissimilar learning processes or sequences in MOOCs, another important aspect worth exploring is the relative frequency of access for each activity type. One way to recognise learners' interests in different learning activities is to analyse the relative frequency of access that also signifies typical learners' experiences within the respective activities (Davis et al. 2017; Liu et al. 2016). In particular, it represents general experience when estimated for an entire cohort. Therefore, this study builds upon the existing literature (Rizvi et al. 2018; Davis et al. 2017; Liu et al. 2016) and aims to explore the linkage (if any) between activity types in a MOOC LD, learners' interests (i.e., expressed through relative frequency of access), and processual learning (i.e., learners' progress in time). In current study, we have investigated and compared the most dominant progression and activity access frequencies within aforementioned three groups of learners.

Research questions

Drawing upon the previous research of understanding learner engagement and progressions through structured learning activities, this study implements and evaluates a two-step approach to understanding learning processes in the context of one FutureLearn science MOOC. We aim to compare three groups of learners that have been identified in prior research (Rizvi et al. 2018), Markers, Partial Markers, and Non-Markers, whose general behaviour signals distinct inclinations towards certification. The goal of this study is to uncover similarities and differences in the learning paths of these three groups with respect to the learning design of the course. We therefore pose the following research questions:

RQ1 How and to what extent does engagement with different elements of the learning design differ between these three groups of learners?

RQ2 How and to what extent do temporal learning paths (i.e., sequences of learning activities) differ between these three groups of learners?

RQ3 How and to what extent can subgroups of learners be identified within each of these three groups, based on the similarity of sequence of learning activities?

Methodology

Context and data

FutureLearn is the largest MOOC provider in Europe and 4th largest in the world in terms of number of enrolled learners (Shah 2016). Compared to other large MOOC providers, FutureLearn follows a social-constructivist pedagogical style by promoting ‘learning through conversations’ (Ferguson and Clow 2015). The course structure comprises a variety of activities: articles, discussion, peer review, quizzes, tests, videos, audio recordings and exercises. Using the theoretical framework for LD discussed in “MOOC event logs and learning processes” section, the majority of FutureLearn courses have a balance of assimilative, communication, adaptive, and assessment activities. The MOOC structure comprised two types of assimilative activities (Video, Article), two types of assessment activities (Test, Quiz) and one communication activity (Discussion). All step categories were available to learners for free, except Test. The assessment activity Test was only available to ‘upgraded’ learners, i.e., learners who had upgraded a MOOC after paying a certain fee, potentially to obtain unlimited access and a certificate. Unlike Quiz activity, which allowed unlimited attempts, Tests had a maximum of three attempts. Learners’ Test scores were then reported on progress page and certificate transcript.

Data for this study were collected in a science MOOC developed by the Open University, which was offered in year 2017 on the FutureLearn platform. The course enrolled a total of 2086 learners and contained 68 learning activities, offered over a span of 4 weeks. Based on how many activities learners have marked as complete in the course, in line with Rizvi et al. (2018) we grouped the study sample into 449 Markers, 832 Partial-Markers, and 805 Non-Markers. For the purpose of our analysis, we extracted the following information from the log files: anonymised learners ID, week number, learning activity-type, learning activity, and timestamps. After the data were collected, we employed the OULDI framework to map the specific activities to general learning design features. Prior to commencing the study, ethical clearance was sought from Human Research Ethics Committee (HREC) at the Open University (OU).

Data analysis

In order to understand learners’ progression, as highlighted in “MOOC event logs and learning processes” section researchers have been using several methods to analyse massive clickstream data extracted from the MOOCs. Educational Data Mining (EDM) methods usually treat these MOOC learning environments as a blackbox (Slater et al. 2017; Baker and Inventado 2014; Papamitsiou and Economides 2014). Traditional EDM plays with sophisticated, hidden patterns that are typically input/output-centric, and not process-centric (Bogarín et al. 2018; Slater et al. 2017). Therefore, in order to obtain a potentially better understanding of learners’ temporal (time-based) behavioural patterns necessitates constructing learners’ navigational patterns (or navigational events) throughout the learning activities.

In this context, several advanced methods are increasingly being used by other researchers. These advanced methods include, but are not limited to, Natural Language Processing (NLP), Sequential Pattern Mining, or associated Stochastic/Probabilistic predictive methods, such as Hidden Markov Models, and/or illustrative methods, such as Graph Mining or Social Network Analysis (SNA) (Geigle and Zhai 2017; Rizvi and Ghani 2016; Robinson et al. 2016; Wen and Rosé 2014). Sequential Pattern Mining and related methods are suitable for finding partial, subsequent sets of learning events. Similarly, these methods along with SNA provide illustrative results of learning engagement, and are particularly suited to find local processes, short sequences, and subgraphs of interest. Nonetheless, such methods may not be appropriate to understand end-to-end transitions, or other temporal dynamics of learning trajectories within a MOOC. Another main disadvantage of using such methods is a lack of comprehensive understanding of end-to-end learning paths followed by large subgroups of learners (Bogarín et al. 2018; Bannert et al. 2014). Therefore, to develop learners' temporal navigational patterns, this study used methods typically associated with Educational Process Mining (EPM).

Process Mining is a set of emerging techniques aimed at extracting process-related knowledge from the events logs. EPM is an application of Process Mining techniques in the educational domain (Bogarín et al. 2018). Apart from drawing the end-to-end learning processes, EPM methods also assist in the comparison of executed processes with normative/intended models (referred to as conformance checking). In Process Mining, the term *Variant* refers to a simplistic view of end-to-end sequence of activities, followed by significant number of cases. Figure 2 clarifies the concept of this term.

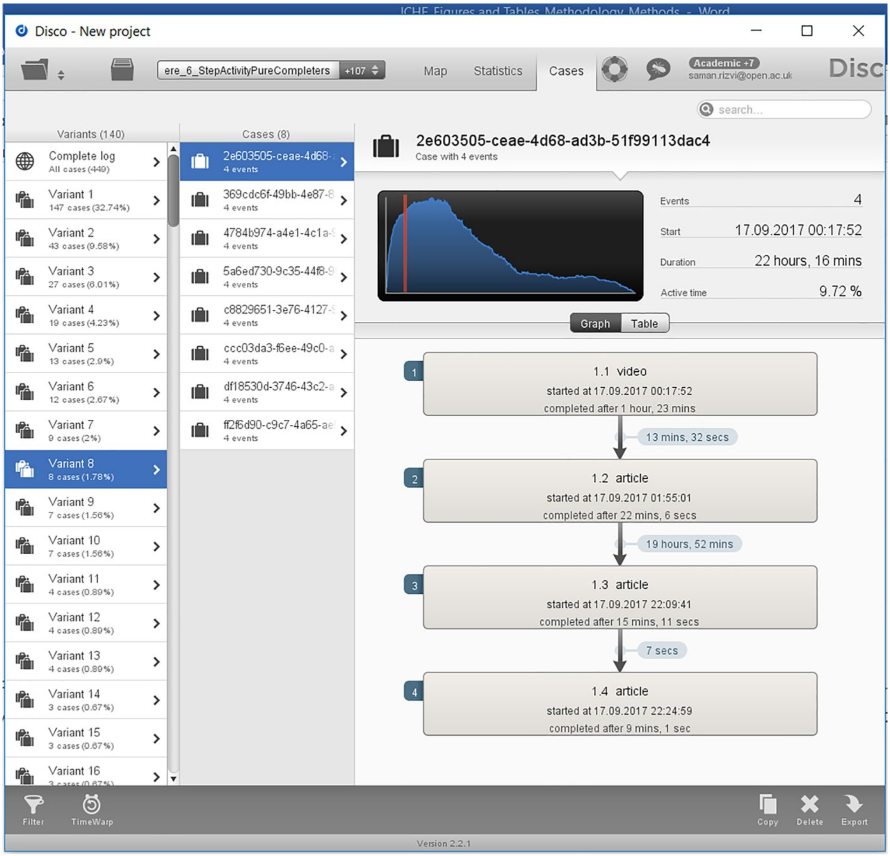

Fig. 2 A list of 140 types of Markers' learning sessions. The type 8 shows 4 end-to-end interactions (events), with the time associated (variant 8: a typical, simplistic learning path of subgroup of 8 Markers)

Our current study focuses on the estimation and comparison of activity access frequency, and temporal learning pathways of dominant subgroups of learners in all three groups. Each of the three groups demonstrated a relatively unique learning process, and all learners from a respective subgroup tended to follow a particular learning pathway in a MOOC. For the construction of process maps, *Discovery* software was used, whereby we used an extended and improved version of *Fuzzy Miner* algorithm (Günther and Van Der Aalst 2007), which creates elaborative, uncomplicated process maps and can easily identify infrequent subgroups. To improve the statistical soundness of our arguments and to see if the subgroups from these three groups were actually different, we used Chi square method.

Results

In the exploratory phase of our analysis, we found three distinct clicking patterns that led us to the learners' categorisation we used in this study; we identified three groups; Markers, Partial-Markers and Non-Markers. The categorisation appeared to be unique within the relevant FutureLearn context, although this categorisation is partially derived from, and partly based upon, similar categorisation used in previous MOOC engagement literature (Davis et al. 2016; Ferguson and Clow 2015; Guo and Reinecke 2014). As can be seen in Fig. 3, the group of Markers remained far more active throughout the MOOC than Partial and Non-Markers in terms of hourly activity. This was particularly noticeable during the first half of the course, whereas overall activity levels diminished with time for all learners afterwards.

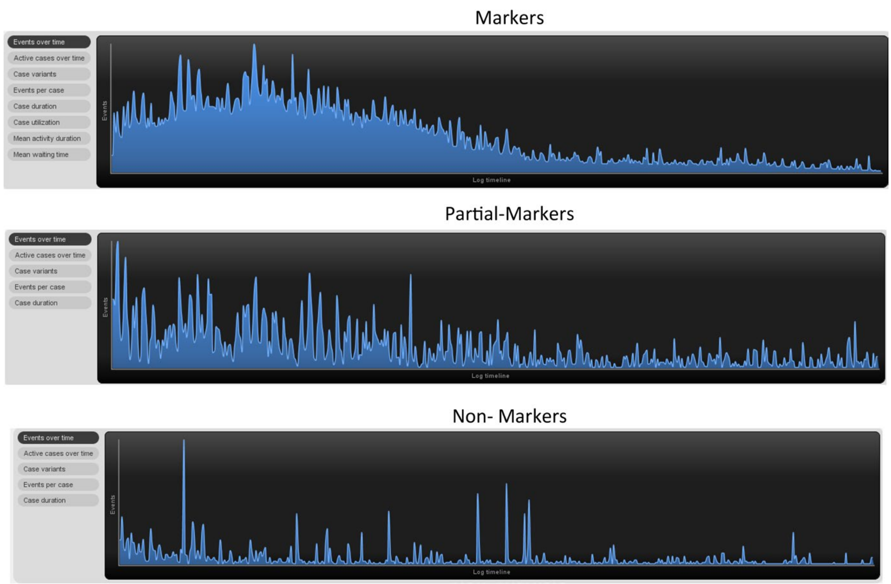

Fig. 3 Difference of engagement behavior in all three groups

In week 1, Markers largely accessed some articles (accessed 3876 time), closely followed by discussion (1135), video (804) and quiz (365). However, typically they spent most time watching video (median up to 8 min 6 s) and spent least time on reading an article (median up to 2 min 48 s). Partial-Markers followed the same pattern. In contrast, Non-Markers preferred watching videos (50% of their overall activities in week 1), followed by article (40.1%), discussion (6.98%), and quiz (2.05%) respectively, but without marking any of the activity as completed in week 1. In week 2, all three groups remained mostly interested in articles. Although discussion was found to be second most frequent activity, learners started to spend less time overall in participating in a discussion (just more than 1 min in case of Markers). In week 3 and 4, Partial- and Non-Markers gradually withdrew from discussions, however they continued to read articles and viewed videos as before. While Markers remained mildly interested in participating in discussion, still typically spending less than 2 min on a discussion activity in last 2 weeks.

RQ1 Variation in engagement with elements of the learning design.

In order to analyse variation in learning behaviour across the three groups, and in line with the prior work of Rizvi et al. (2018), Davis et al. (2017) and Liu et al. (2016), we utilised relative frequency of access for each activity type in relation to the activity distribution in the MOOC. As discussed in "Literature review" section, the relative access frequency can be representative of learners' interests, or a wish to engage with a particular activity type. Furthermore, relative frequency of access also represents (part of the) general experience of the entire cohort.

Figure 4 illustrates the distribution of engagement with course activities for the three groups (raw frequencies are provided in "Appendix" Table 2). We found that while Markers and Partial-Markers engagement in assimilative and communication activities is equivalent, Markers are more engaged in assessment activities than Partial-Markers. In contrast, Non-Markers were most engaged with a specific assimilative activity, video watching, but less engaged in other assimilative and communication activities: reading articles and participating in discussion. Non-Markers were also notably less engaged in assessment activities compared to Markers and Partial-Markers. This may be attributed to Non-Markers' lack of interest in active participation or certification attainment.

RQ2 Variation in temporal learning paths.

In order to address RQ2 and RQ3 we mapped the learning paths based on the clickstream data and identified main subgroups within each group. Omitting the self-loop (i.e. repetition) provided more clarity to the process maps. For example, Fig. 5 shows a simplified view of the learning process model for Markers, filtering out some less frequently occurring pathways. Activity access frequency is also denoted alongside each path.

A closer inspection of end-to-end learning pathways confirmed that although a main pathway existed (dark, thick lines on the map), a large number of Markers preferred non-linear, highly unstructured pathways through the course content. For example, Fig. 5 shows 22 Markers skipping an assimilative activity (Article:

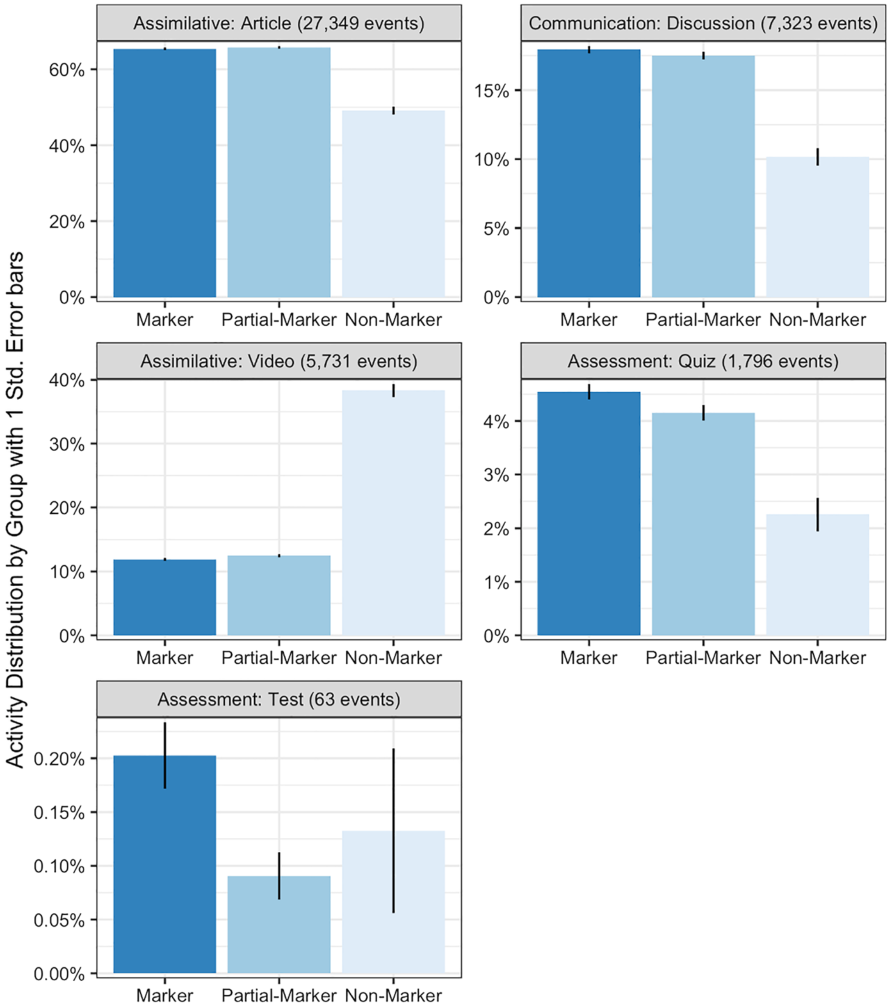

Fig. 4 Distribution of engagement with course activities as classified by the learning design taxonomy among Markers, Partial-Markers, and Non-Markers. Error bars represent 1 standard error

Activity 1.6) to participate in the subsequent activity (Activity 1.7) which was discussion-based. This non-linear progression was consistently noticed in all three groups but, counter intuitively, persisted mainly in Markers.

RQ3 Subgroups identification.

We compared the 15 most common subgroups identified within each of the three primary groups (data available in "Appendix" Table 3). These 15 subgroups account for different amounts of the overall variance in each group: 68.6% for Markers, 46.5% for Partial-Markers, and 89.8% for Non-Markers. This distribution shows that there was more variance in the learning processes among Partial-Markers than the other two groups of

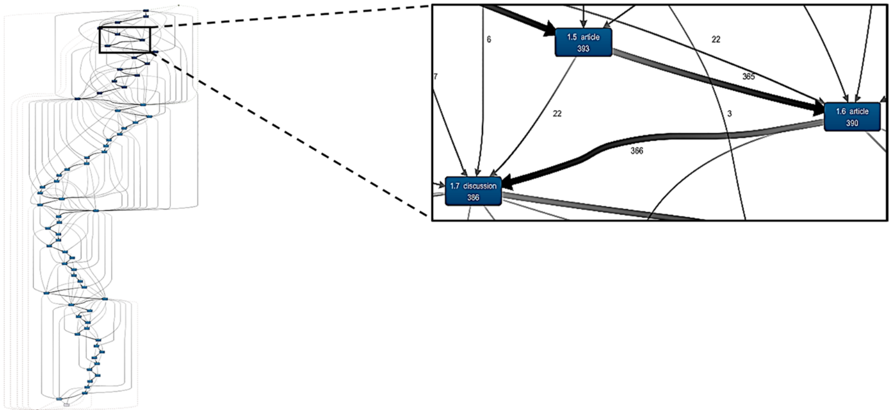

Fig. 5 A simplified view of Markers learning process

learners, because their overall behaviour was captured less accurately by a small number of subgroups. For each subgroup, we computed the number of activities contained in the learning process. We found that a third of Markers (31.4%) followed a long learning process containing 67 distinct activities. In contrast, two thirds of Non-Markers (67.7%) followed a learning process that only contained one activity before they dropped out of the course. In keeping with this pattern, we found that among the top 15 subgroups, Markers tended to have longer learning processes (6 out of 15 with 50 or more activities), Non-Markers had only short learning processes (11 out of 15 with 5 or fewer activities), and Partial-Markers exhibited a mixture of shorter and longer learning processes (2 out of 15 with 50 or more activities; 4 out of 15 with 5 or fewer activities).

To test the robustness of the observed pattern of variation, we performed a set of χ^2 tests of independence. The results indicated that there was a significant association between type of learning activity and whether learner was a Marker, Partial-Marker or Non-Marker ($\chi^2=1279$, $df=8$, $p<0.001$). We also confirmed that the lengths of the learning processes were significantly different across the three groups ($\chi^2=523$, $df=28$, $p<0.001$).

Discussion and conclusion

The purpose of this exploratory study was to determine the nature and extent of differences in participatory behaviour and temporal learning paths of MOOC learners, in the light of learning activity type attributed from an established learning design model. Another aim of this investigation was to understand the common pathways followed by a substantially large subgroup of learners, referred to as variants in process mining. We found the progression trend for individual groups remained aligned with our previous work (Rizvi et al. 2018) and with other MOOC literature (Kizilcec et al. 2013; Ferguson and Clow 2015). Our current study employed an established learning design taxonomy to investigate the detailed processes of engagement over time. This study extends our prior work that has identified three primary clusters of engagement in courses and uncovered similarities and differences in the learning paths of these three groups with respect to the learning design of the course.

Notwithstanding the distinct patterns of engagement with different type of activities, the results remained very similar to previous studies in formal online learning setting (Nguyen et al. 2017) showing an overall liking of assimilative activities in general and video-based assimilative activities in particular. Taken together, these results provide insights into learners' temporal progression or pathways in the MOOC. Our overall findings are aligned with the previous research in MOOC learning environment (Ferguson and Clow 2015). While we noticed that top subgroups in all groups left the MOOC right after accessing an assimilative activity (either video or article), and very rarely after accessing an assessment activity or participating in a discussion.

The findings also suggest that academics and course designers should give more thought into designing communication or assessment activities for MOOC learning environment, in order make to such activities more appealing to an informal learner. The findings from this study suggested that Markers and Partial-Markers access frequencies for all activity types were found to be either aligned with the MOOC distribution or else exceeded expectations. Non-Markers demonstrated huge early drop-outs, however if they continued they remained substantially interested in assimilative activities of video watching. This result points that in general, Non-Markers remained interested in video-based content, and not in the textual content per se (whether assimilative or communicative).

We found substantially large number of learners, from all groups, dropping out after participating in one of the assimilative activities. Since the activity engagement behaviour differed in all three groups of learners, we can deduce that that if analyses were done without categorising the learners, the results would have remained strongly biased towards majority class (Partial-Markers in this case). This suggests that while investigating the temporal and engagement behaviour of learners, it is necessary to first categorise the learners into natural groups.

The study contributes to the field by interrogating the behaviour of learners, while considering different categories that go beyond simply looking at those who completed a substantial fraction of the course, or those who dropped out. This leaves a door open to further research on learners' experiences. i.e. while navigating the course, how are they making these decisions to engage more with one or the other type of activity. As mentioned elsewhere, success in MOOCs is relative, still, without a deep knowledge of learners' navigation through the system, it would remain hard to distinguish between good decisions and bad decisions.

The findings from this study can be beneficial for practice in MOOC learning design and are suggestive of the fact that analyses of voluminous data being captured and stored in MOOC clickstream logs, require innovative methods, such as process mining and variant mapping. Such methods intrinsically support exploration of learners' behaviour hidden in voluminous data. Despite its exploratory nature, current study lays the ground work for our future research into behavioural modelling and mapping within MOOC learning environment. In future, more contextual information or demographic data would help us to establish a greater degree of accuracy on this matter.

Acknowledgements This study is intended for publication in the Special Issue on Current Trends in E-learning Assessment. A previous version of this paper was Rizvi et al. (2018).

Appendix

See Tables 2 and 3.

Table 2 Frequency and relative frequency of access for individual activity type

Activity type	Expected rel. freq. of access	Markers		Partial-Markers		Non-Markers	
	Activity distribution	Frequency	Rel. freq. of access (%)	Frequency	Rel. freq. of access (%)	Frequency	Rel. freq. of access (%)
Assimilative_Article	44 (64.7%)	13,883	65.42	12,354	65.79	1112	49.14
Communication_Discussion	12 (17.6%)	3806	17.94	3287	17.5	230	10.16
Assimilative_Video	8 (11.7%)	2523	11.89	2341	12.47	867	38.31
Assessment_Quiz	3 (4.4%)	965	4.55	780	4.15	51	2.25
Assessment_Test	1 (1.5%)	43	0.2	17	0.09	3	0.13

Table 3 Most common subgroups within three primary group of learners

Subgroups	Markers		Partial-Markers		Non-Markers	
	Cases (449)	Events	Cases (832)	Events	Cases (805)	Events
V1	141 (31.4%)	67	73 (8.77%)	2	545 (67.7%)	1
V2	44 (9.8%)	16	44 (5.29%	3	80 (9.94%)	2
V3	28 (6.24%)	1	31 (3.73%)	4	28 (3.48%)	3
V4	19 (4.23%)	68	31 (3.73%)	6	23 (2.86%)	1
V5	13 (2.9%)	34	28 (3.37%)	7	8 (0.99%)	2
V6	11 (2.45%)	50	25 (3%)	5	7 (0.87%)	4
V7	10 (2.23%)	65	23 (2.76%)	16	6 (0.75%)	5
V8	8 (1.78%)	4	23 (2.76%)	67	5 (0.62%)	1
V9	7 (1.56%)	3	22 (2.64%)	8	4 (0.5%)	6
V10	7 (1.56%)	68	21 (2.52%)	9	4 (0.5%)	3
V11	4 (0.89%)	2	14 (1.68%)	10	3 (0.37%)	14
V12	4 (0.89%)	6	14 (1.68%)	66	3 (0.37%)	2
V13	4 (0.89%)	8	12 (1.44%)	11	3 (0.37%)	2
V14	4 (0.89%)	9	12 (1.44%)	34	2 (0.25%)	7
V15	4 (0.89%)	67	11 (1.32%)	13	2 (0.25%)	9

References

Baker, R., Evans, B., & Dee, T. (2016). A randomized experiment testing the efficacy of a scheduling nudge in a massive open online course (MOOC). *AERA Open, 2*(4), 2332858416674007.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61–75). Retrieved from http://link.springer.com/10.1007/978-1-4614-3305-7_4.

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning, 9*(2), 161–185.

Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(1), e1230.

Conole, G. (2012). *Designing for learning in an open world* (Vol. 4). Berlin: Springer.

Cross, S., & Conole, G. (2009). *Learn about learning design. Part of the OU Learn about series of guides*. The Open University: Milton Keynes. Retrieved from http://www.open.ac.uk/blogs/OULDI/wp-content/uploads/2010/11/Learn-about-learning-design_v7.doc.

Cross, S., Galley, R., Brasher, A., & Weller, M. (2012). *OULDI-JISC project evaluation report: The impact of new curriculum design tools and approaches on institutional process and design cultures*.

Davis, D., Chen, G., Hauff, C., & Houben, G.-J. (2016). Gauging MOOC learners' adherence to the designed learning path. In *9th international conference on EDM*.

Davis, D., Jivet, I., Kizilcec, R. F., Chen, G., Hauff, C., & Houben, G.-J. (2017). Follow the successful crowd: raising MOOC completion rates through social comparison at scale. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 454–463). ACM.

Davis, D., Seaton, D., Hauff, C., & Houben, G.-J. (2018). Toward large-scale learning design.

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). ACM.

Geigle, C., & Zhai, C. (2017). Modeling MOOC student behavior with two-layer hidden Markov models. In *Proceedings of the fourth (2017) ACM conference on learning@ scale* (pp. 205–208). ACM.

Günther, C. W., & Van Der Aalst, W. M. (2007). Fuzzy mining—Adaptive process simplification based on multi-perspective metrics. In *International conference on business process management* (pp. 328–343). Berlin: Springer.

Guo, P. J., & Reinecke, K. (2014). Demographic differences in how students navigate through MOOCs. In *Proceedings of the first ACM conference on learning@ scale conference* (pp. 21–30). Retrieved from http://dl.acm.org/citation.cfm?id=2566247.

Henderikx, M., Kreijns, K., & Kalz, M. (2017). An alternative approach for measuring MOOC success based on participant's intentions.

Jansen, D., & Schuwer, R. (2015). *Institutional MOOC strategies in Europe. Status report based on a mapping survey conducted in October–December 2014*. EADTU.

Joksimović, S., Poquet, O., Kovanović, V., Dowell, N., Mills, C., Gašević, D., et al. (2017). How do we model learning at scale? A systematic review of research on MOOCs. *Review of Educational Research, 88,* 43–86.

Juhaňák, L., Zounek, J., & Rohlíková, L. (2017). Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. *Computers in Human Behavior, 92,* 496–506.

Kizilcec, R. F., Davis, G. M., & Cohen, G. L. (2017). Towards equal opportunities in MOOCs: affirmation reduces gender & social-class achievement gaps in China. In *Proceedings of the fourth (2017) ACM conference on learning@ scale* (pp. 121–130). ACM.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 170–179). ACM.

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

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.

Liu, Z., Brown, R., Lynch, C., Barnes, T., Baker, R. S., Bergner, Y., et al. (2016). MOOC learner behaviors by country and culture; an exploratory analysis. *EDM, 16,* 127–134.

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Munoz-Gama, J. (2018). Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in massive open online courses. *Computers in Human Behavior, 80,* 179–196.

Mangaroska, K., & Giannakos, M. N. (2018). Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. *IEEE Transactions on Learning Technologies*. https://doi.org/10.1109/TLT.2018.2868673.

Margaryan, A., Bianco, M., & Littlejohn, A. (2015). Instructional quality of massive open online courses (MOOCs). *Computers & Education, 80,* 77–83.

Mayer, R. E. (2005). *The Cambridge handbook of multimedia learning*. Cambridge: Cambridge University Press.

Milligan, C., & Littlejohn, A. (2017). Why study on a MOOC? The motives of students and professionals. *The International Review of Research in Open and Distributed Learning, 18*(2), 92–102.

Morgado, L., Mota, J., Jansen, D., Fano, S., Tomasini, A., Silva, A., Fueyo Gutiérrez, A., Giannatelli, A., Brouns, F. (2014). ECO D2. 2 Instructional design and scenarios for MOOCs version 1.

Nguyen, Q. (2017). Unravelling the dynamics of learning design within and between disciplines in higher education using learning analytics.

Nguyen, Q., Huptych, M., & Rienties, B. (2018). Linking students' timing of engagement to learning design and academic performance. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 141–150). ACM.

Nguyen, Q., Rienties, B., & Toetenel, L. (2017). Mixing and matching learning design and learning analytics. In *International conference on learning and collaboration technologies* (pp. 302–316). Berlin: Springer.

Papamitsiou, Z. K., & Economides, A. A. (2014). Learning Analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society, 17*(4), 49–64.

Rai, L., & Chunrao, D. (2016). Influencing factors of success and failure in MOOC and general analysis of learner behavior. *International Journal of Information and Education Technology, 6*(4), 262.

Rienties, B., & Toetenel, L. (2016). The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules. *Computers in Human Behavior*, *60*, 333–341.

Rizvi, S., & Ghani, S. (2016). Predicting higher education MOOCs engagement-level odds; a stochastic approach. In *Presented at the society for research in higher education international annual research conference (SRHE-16)*. Newport, United Kingdom.

Rizvi, S., Rienties, B., & Khoja, S. A. (2019). The role of demographics in online learning; A decision tree based approach. *Computers & Education*, *137*, 32–47.

Rizvi, S., Rienties, B., & Rogaten, J. (2018). Temporal dynamics of MOOC learning trajectories. In *Proceedings of the international conference on data science, E-learning and information systems*. DATA'18. https://doi.org/10.1145/3279996.3280035.

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., & Gehlbach, H. (2016). Forecasting student achievement in MOOCs with natural language processing. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 383–387). Retrieved from http://dl.acm.org/citation.cfm?id=2883932.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12–27.

Selwyn, N. (2015). Data entry: Towards the critical study of digital data and education. *Learning, Media and Technology, 40*(1), 64–82.

Sergis, S., Sampson, D. G., & Pelliccione, L. (2017). Educational design for MOOCs: Design considerations for technology-supported learning at large scale. In *Open education: From OERs to MOOCs* (pp. 39–71). Berlin: Springer.

Shah, D. (2016). Monetization over massiveness: Breaking down MOOCs by the numbers in 2016. *EdSurge*. https://www.Edsurge.Com/. Accessed July 25, 2017.

Sharples, M. (2015). FutureLearn learning design guidelines.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics, 42*(1), 85–106.

Sonnenberg, C., & Bannert, M. (2015). Discovering the effects of metacognitive prompts on the sequential structure of SRL-processes using process mining techniques. *Journal of Learning Analytics, 2*(1), 72–100.

Sparke, M. (2017). Situated cyborg knowledge in not so borderless online global education: Mapping the geosocial landscape of a MOOC. *Geopolitics, 22*(1), 51–72.

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains. *International Educational Data Mining Society*.

Wen, M., & Rosé, C. P. (2014). Identifying latent study habits by mining learner behavior patterns in massive open online courses. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management* (pp. 1983–1986). ACM.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81.

Winne, P. H. (2017). Leveraging big data to help each learner and accelerate learning science. *Teachers College Record, 119*(3), 1–24.

Saman Rizvi is a PhD candidate at Institute of Educational Technology, The Open University. Her PhD research focuses on the role of geocultural background in MOOC learning and is fully funded by the Leverhulme Trust for Open World Learning project. She has an academic background in computer science and information technology. She holds a researcher and risk-analyst level expertise in data mining, statistical and stochastic data modelling and is familiar with a variety of machine learning and statistical modelling tools. She also has experience in e-Learning development, training, support, and front- and back-end administration of enterprise-scale Learning Management Systems. Her research interests include but are not limited to socio-cultural influences on online learning, Massive Open Online Courses (MOOCs), inclusive and diverse online learning environment, computer-supported collaborative learning.

Dr. Bart Rienties is Professor of Learning Analytics at the Institute of Educational Technology at the Open University UK. He is programme director Learning Analytics within IET and head of Data Wranglers, whereby he leads of group of learning analytics academics who conduct evidence-based research and sense making of Big Data at the OU. As educational psychologist, he conducts multi-disciplinary research on work-based and collaborative learning environments and focuses on the role of social interaction in learning, which is published in leading academic journals and books. His primary research interests are focussed on Learning Analytics, Computer-Supported Collaborative Learning, and the role of motivation in learning. Furthermore, Bart is interested in broader internationalisation aspects of higher education. He has successfully led a range of institutional/national/European projects and received a range of awards for his educational innovation projects.

Dr. Jekaterina Rogaten is a senior lecturer and course Leader for MSc Applied Psychology in Fashion at London College of Fashion, University of the Arts London. Her main research interests are in the field of positive psychology and education. In particular, she has expertise in the evaluation of interventions and teaching within Higher Education. Most of her research is cantered but not limited to evaluation of the effectiveness of Higher Education courses and programmes as well as the effect of individual differences of academic progression. In addition, she is also interested in the evaluation of online learning environments and assessments of the learning design impact on learning. She also conducts research into the effects of psychological and socio-demographic predisposition on students' learning and development of skills and abilities.

René F. Kizilcec is an Assistant Professor in the School of Computing and Information Science at Cornell University, where he directs the Future of Learning Lab. His research is on the impact of digital technologies in formal and informal learning contexts and scalable interventions to broaden participation, raise academic performance, and reduce achievement gaps. He leverages techniques from data mining, machine learning, and natural language processing to examine behavior and motivation, reveal heterogeneous treatment effects, and inform user-centered design. Kizilcec received a BA in Philosophy and Economics from University College London, and an MSc in Statistics and PhD in Communication from

Stanford, with a thesis on designing psychologically welcoming online learning environments, which was awarded the Nathan Maccoby Outstanding Dissertation Award. Prior to joining Cornell, he was a research scientist in Facebook's Core Data Science team, research director in the Stanford Graduate School of Education, and research professor at Arizona State University.

Affiliations

Saman Rizvi[1] · Bart Rienties[1] · Jekaterina Rogaten[2] · René F. Kizilcec[3]

Bart Rienties
bart.rienties@open.ac.uk

Jekaterina Rogaten
j.rogaten@fashion.arts.ac.uk

René F. Kizilcec
kizilcec@cornell.edu

[1] Institute of Educational Technology, The Open University, Milton Keynes, UK

[2] University of the Arts London, London, UK

[3] Cornell University Ithaca, Ithaca, NY, USA