Elsevier Editorial System(tm) for Computers in Human Behavior

Manuscript Draft

Manuscript Number:

Title: Mining Theory-Based Patterns from Big Data: Identifying Self-Regulated Learning Strategies in Massive Open Online Courses

Article Type: Full Length Article

Section/Category: Full Length Article

Keywords: Massive Open Online Courses; Interaction Sequence Patterns; Self-Regulation; Process Mining; Learning Sequences; Big data

Corresponding Author: Mr. Jorge Javier Maldonado Mahauad, Mgti, Mtiae

Corresponding Author's Institution: Universidad de Cuenca / Pontificia Universidad Católica de Chile

First Author: Jorge Javier Maldonado Mahauad, Mgti, Mtiae

Order of Authors: Jorge Javier Maldonado Mahauad, Mgti, Mtiae; Mar Pérez-Sanagustín, Ph.D; René F Kizilcec, Ph.D (c); Nicolás Morales; Jorge Munoz-Gama, Ph.D

Abstract: Big data in education offers unprecedented opportunities to support learners and advance research in the learning sciences. Analysis of observed behaviour using computational methods can uncover patterns that reflect theoretically established processes, such as those involved in self-regulated learning (SRL). This research addresses the question of how to integrate this bottom-up approach of mining behavioural patterns with the traditional top-down approach of using validated self-reporting instruments. Using process mining, we extracted interaction sequences from fine-grained behavioural traces for 3,458 learners across three Massive Open Online Courses. We identified seven common distinct interaction sequence patterns. High-SRL learners were more likely to watch multiple video-lectures and solve multiple assessments in a sequence than low-SRL learners. Specifically, high-SRL learners who completed the course were more likely to watch video-lectures before passing an assessment than low-SRL learners who completed the course. By contrast, low-SRL completers were more likely to take assessments instead of watching video-lectures. We discuss challenges that arose in the process of extracting theory-based patterns from observed behaviour, including analytic issues and limitations of available trace data from learning platforms. Harnessing learners' detailed behavioural records, unlike questionnaire data, can provide an objective longitudinal account of learning and enable real-time support and feedback.

Highlights

- Self-regulated learning (SRL) is necessary for learners in MOOCs to succeed
- MOOC data helps understanding theoretical models in SRL
- We integrated bottom-up approach of mining behavioural patterns with traditional self-reported instruments
- We identified seven distinct learners' interaction sequences with the MOOC content
- We also found differences between sequences of low- and high SRL learners

*Title Page

Mining Theory-Based Patterns from Big Data: Identifying Self-Regulated Learning Strategies in Massive Open Online Courses

Jorge J. Maldonado[a,b,*], Mar Pérez-Sanagustín[a], René F. Kizilcec[c], Nicolás Morales[a], Jorge Munoz-Gama[a]

[a]Pontificia Universidad Católica de Chile, Department of Computer Science, Chile
[b]Universidad de Cuenca, Department of Computer Science, Ecuador
[c] Stanford University, Department of Communication, USA
[*]Corresponding autor

Email addresses: jjmaldonado@uc.cl, mar.perez@uc.cl, kizilcec@stanford.edu, nvmorale@uc.cl, jmun@uc.cl

Present Addresses:

Jorge J. Maldonado
Department of Computer Science
Pontificia Universidad Católica de Chile
Avda. Vicuña Mackenna 4860, Macul
Santiago
Chile

Mar Pérez-Sanagustín
Department of Computer Science
Pontificia Universidad Católica de Chile
Avda. Vicuña Mackenna 4860, Macul
Santiago
Chile

René F. Kizilcec
Department of Communication
450 Serra Mall
Stanford University
Stanford, CA 94305
USA

Nicolás Morales
Department of Computer Science
Pontificia Universidad Católica de Chile
Avda. Vicuña Mackenna 4860, Macul
Santiago
Chile

Jorge Munoz-Gama
Department of Computer Science
Pontificia Universidad Católica de Chile
Avda. Vicuña Mackenna 4860, Macul
Santiago
Chile

Compliance with Ethical Standards: informed consent was obtained for use survey-instrument with human subjects.

Conflict of Interest: The authors declare that they have no conflict of interest.

Mining Theory-Based Patterns from Big Data: Identifying Self-Regulated Learning Strategies in Massive Open Online Courses

(Author names and affiliations omitted for blind review)

Abstract

Big data in education offers unprecedented opportunities to support learners and advance research in the learning sciences. Analysis of observed behaviour using computational methods can uncover patterns that reflect theoretically established processes, such as those involved in self-regulated learning (SRL). This research addresses the question of how to integrate this bottom-up approach of mining behavioural patterns with the traditional top-down approach of using validated self-reporting instruments. Using process mining, we extracted interaction sequences from fine-grained behavioural traces for 3,458 learners across three Massive Open Online Courses. We identified seven common distinct interaction sequence patterns. High-SRL learners were more likely to watch multiple video-lectures and solve multiple assessments in a sequence than low-SRL learners. Specifically, high-SRL learners who completed the course were more likely to watch video-lectures before passing an assessment than low-SRL learners who completed the course. By contrast, low-SRL completers were more likely to take assessments instead of watching video-lectures. We discuss challenges that arose in the process of extracting theory-based patterns from observed behaviour, including analytic issues and limitations of available trace data from learning platforms. Harnessing learners' detailed behavioural records, unlike questionnaire data, can provide an objective longitudinal account of learning and enable real-time support and feedback.

Keywords: Massive Open Online Courses, Interaction Sequence Patterns, Self-Regulation, Process Mining, Learning Sequences, Big data

1.- Introduction

In recent years, masses of fine-grained educational records became available to researchers and fuelled the nascent science of learning analytics (Dietze, Siemens, Taibi, & Drachsler, 2016). Digital learning platforms such as Massive Open Online Courses (MOOCs) collect detailed records of each learner's behaviour, performance, and other types of interaction. Nevertheless, despite the large amount of data that MOOCs are collecting, this information may not be sufficient to understand theory if it is not processed and analysed carefully. In fact, as Lodge & Lewis (2012) state, we only have access to certain limited kinds of data, and this data does not necessarily offer valuable information about learners' behaviour and learning processes. Large amounts of data allow us to extract patterns about what learners do through data-driven methods, but these methods are not enough to understand more deeply what these patterns mean and how they relate with theory. Furthermore, the data collected become more valuable when combining the data from platforms (capturing the learners' actual interactions) with learners' self-reports (e.g. through surveys that allow us to capture what they say they do) (Eynon, 2013). Therefore, there is a need to explore new ways to connect data-driven methods with learners' self-reported data and theory to get a better understanding of how learners behave and learn in digital environments (Lodge & Corrin, 2017).

Nowadays, MOOC platforms reach thousands of learners worldwide (Breslow et al., 2013; Cooper and Sahami, 2013; Daradoumis et al., 2013), becoming one of the biggest sources of learners' recorded behaviour. They can be used to gain new knowledge about how learners behave in online environments. In this paper, we use MOOC data to advance the research of self-regulated learning (SRL) online. Recent studies show that in order for MOOC learners to achieve their objectives, they must have the capacity to regulate their own learning (Hew & Cheung, 2014; Kizilcec & Schneider, 2015). Self-regulated learners are characterised by their ability to initiate cognitive, metacognitive, affective and motivational processes (Boekaerts, 1997). Moreover, SRL research indicates that successful learning is associated with the active deployment of regulatory activities during the learning process, such as goal-setting, planning or monitoring (Bannert, 2009; Johnson, Azevedo, & D'Mello, 2011). The ability to develop

these strategies is an essential skill in order to succeed in an open context such as a MOOC, where the learner should advance independently without support from a tutor or professor. However, how people self-regulate in a MOOC is still an open question.

In the last 30 years, a large number of models have been developed that explain how the process of SRL develops amongst learners (Boekaerts, 1999; Borkowski, 1996; Pintrich, 2004; Winne & Hadwin, 1998; Zimmerman, 2015). These models served as a foundation for developing methods to study the use of SRL strategies in the learning process. They can be categorised as component models and process models (Wirth & Leutner, 2008). Component models describe SRL in terms of different strategies that promote or encourage self-regulation, which are seen as long-lasting characteristics of a person. These models describe self-regulation strategies independently of the stage in the learning process at which they are necessary. Examples of these models are those developed by Boekaerts (1999) and Pintrich (2000). In comparison, process models can describe the "ideal" SRL process as a series of phases that occur during the learning process. Process models describe typical requirements that learners have to meet in different phases of the cyclical learning process, but they do not specify the strategies necessary to meet those requirements (Zimmerman, 1998). Examples of these models are those developed by Zimmerman (2000), Borkowski, (1996), and Winne & Hadwin (1998). Depending on the model that is used as a reference, SRL in a MOOC can be studied from these two perspectives: either as an aptitude if component models are used, or as a process if process models are used (Winne, 2010).

In online environments, the most common approach is to study SRL as an aptitude. Many instruments have been developed over the last decade to measure which SRL strategies learners use in online environments, including think-aloud protocols (a type of interview) and learning diaries (Roth, Ogrin, & Schmitz, 2015; Wirth & Leutner, 2008). Yet self-report questionnaires are the most common type of assessment of learners' SRL profiles (Roth et al., 2015). Recent studies have adapted and applied questionnaires to determine the level of self-regulation among MOOC learners (Beheshitha et al., 2015; Littlejohn et al., 2016; Kizilcec et al., 2016; Jansen et al., 2016). These studies offer an account of learners' self-regulation profiles in a MOOC and how they relate to their results in the course using a variety of instruments and methods (e.g. clustering, penalized regression). However, none of them have analysed SRL from a process perspective.

The study of the SRL in online environments as a process has gained attention from researchers in the past several years. Researchers have moved from an aptitude-oriented approach to a process-oriented approach. Since SRL can be conceived as a set of events or actions that learners perform (as a process), rather than descriptions of those actions or mental states that these actions generate (Bannert, Reimann, & Sonnenberg, 2014), Process Mining (PM) is a suitable approach for studying SRL in online environments. PM is ideally suited for the analysis of behaviour from a process perspective. In particular, PM facilitates the discovery of learning process models, which represent the workflow of learners' interactions with course materials (van Der Aalst et al., 2011). PM also provides robust ways of extracting, analysing and visualising learners' interaction traces (Mukala, Buijs, & van Der Aalst, 2015b; Romero et al., 2016; Jivet, 2016). These interaction traces are temporal sequences of events of learners' behaviour in the online environment that allow tracing of aptitudes in natural settings (Winne, 2014). Researchers have developed controlled online environments in order to study SRL as a process. For example, Hadwin et al. (2007) examined the performance of eight learners across two study sessions on the gStudy platform. They compared traces of actual study activities to self-reporting on SRL and found that students' self-reports may not align with actual studying activity. More recently, Beheshitha et al. (2015) examined the relationships between 22 undergraduate learners' self-reported SRL aptitudes—such as achievement goal orientation and learning approaches—and the strategies they followed in a learning environment on the nStudy tool. They found differences in transitions between the SRL cognitive strategies performed by both deep and surface learners. Sonnenberg and Bannert (2015) analysed sequential patterns in the learning process of learners in an online environment. They found that using metacognitive prompts to support learners' SRL had an effect on the order in which they participated in learning activities. In a recent experiment in an online environment designed to support SRL at the workplace, Siadaty, Gašević, and Hatala (2016) analysed trace data to build a transition graph of learning actions. The results show that promoting social awareness had the highest correspondence with the SRL processes of the learners.

In MOOCs, PM is becoming an increasingly common technique that provides important mechanisms for understanding learning processes from learners' activity trails obtained from MOOC platform logs (Mukala et al., 2015a). This approach has been used in other online learning environment studies, but with a restriction on the number of participants in the study (generally samples under 30 participants and controlled environments) and homogeneity of the learners. A MOOC environment provides a considerable sample size and adds the heterogeneity

of the learners in an uncontrolled environment, which allows us to extend previous findings. For example, Mukala, Buijs, & van Der Aalst (2015a) applied PM in order to understand learning processes based on learners' interaction in a MOOC with 43,218 learners. The goal was to produce insights to improve the quality of the course (Mukala, Buijs, & van Der Aalst, 2015a). In another study, the same authors used PM techniques to analyse learners' learning patterns in MOOCs. This analysis showed that (1) successful students perform better because they follow the videos and submit quizzes in a more structured way than unsuccessful students; and that (2) regularly watching successive videos in batches had a positive impact on learners' final grades, and a correlation with the interval of time between successive videos they watched (Mukala et al., 2015b).

These studies show the advantages of using PM techniques to understand the behaviour of MOOC learners and relate them to learners' success in courses. However, as a recent study by the MOOC Research Institute indicates, such research is still scarce (Gasevic, Kovanovic, Joksimovic, & Siemens, 2014). Moreover, in the context of SRL and MOOCs, there are still no studies that use PM techniques to learn more about how learners with different SRL profiles perform in courses in terms of their learning sequences.

In order to make progress in this area, the conceptualization of these two approaches to learning (as an aptitude and as a process) makes suitable for study the patterns in observed learner behaviour that reflect theoretically established processes in SRL. In a recent article, we worked on a first approach toward this aim (REFERENCE REMOVED FOR BLIND REVIEW). In this research, we offered an analysis of the relationship between self-reported SRL and actual behaviour in six MOOCs. We found that learners who reported engaging in more SRL behaviour were more likely to achieve their course goals (e.g. completion) and they were more likely to review course materials that they had studied in the past (e.g. reviewing previously attempted assessments). However, in this prior work we studied behaviour at the level of individual interaction (using transition probability to pass from one interaction to another) to obtain a basic process model. However, a deeper approach that considers more complex sequences is needed to understand SRL in MOOCs as a process.

In this paper we extend our previous findings using formal PM techniques in order to go deeper (looking for broad interaction sequences) and understand the relationship between theoretical self-reported SRL strategies and behavioural patterns on large-scale MOOC platforms. Prior work suggests that learners interact with video-lectures, assessments and other MOOC contents week by week, identifying loopbacks, deviations and bottlenecks. We also provide insights in terms of students' learning and assessment submissions behaviour at a high grain scale. However, we found no evidence on how these student activity trails are related with SRL strategies. We therefore pose the following three research questions:

- **RQ1.** What are the most frequent interaction sequences of learner behaviour in MOOCs?
- **RQ2.** Is there any difference in the interaction sequences of learner behaviour between those who complete a course and those who do not?
- **RQ3.** Do interaction sequences of learner behaviour differ between learners with a high-SRL versus a low-SRL profile?

An analysis of the learners' sequences behaviour in a MOOC from a PM perspective will allow us to advance our understanding of SRL in MOOCs, providing insights about how observed interaction sequence patterns are aligned with SRL strategies. To address these research questions, we present the results of an exploratory study carried out in three Coursera courses[1]. The results show that watching video-lectures and solving assessment exercises was significantly more common among learners with a high-SRL than a low-SRL profile. In the following sections, we describe the context of the study, the instruments employed and the PM techniques used for the analysis.

2. Method

This section presents the exploratory study we performed in three MOOCs to address the research questions. Specifically, we present the characteristics of the learners' sample (section 2.1. Sample) and the courses analysed in the study (section 2.2. Courses). We also present the self-reported questionnaire used as an instrument to

[1] Coursera courses: Aula constructivista, Electrones en acción and Gestión de organizaciones.

identify the learners' SRL profiles (section 2.3. Measures), and the procedure developed for analysing the Coursera learners' traces data using PM techniques (section 2.4. Procedure).

2.1. Sample

The final study sample included N = 3,458 online learners in three different MOOCs. This sample was a subset of 4,871 learners who answered the initial questionnaire from the total of 54,935 that registered for the MOOCs. 1,413 responses were removed for various reasons (e.g. questionnaires taken more than once, incomplete answers in the questionnaire). The target audiences of these courses were high school students, college students and professionals in subject-related industries. Based on the demographic data captured during the registration process on the platform, the average age was 32.0 (SD = 11.07). One quarter of learners were women and 88% held a bachelor's degree or higher (14% a master's or Ph.D.). Data collection occurred between August 2015 and June 2016.

2.2. Courses

This study encompassed three courses offered by Pontificia Universidad Católica de Chile on Coursera. The courses were taught in Spanish on topics related to engineering (N = 2,035), education (N = 497) and management (N = 926). The course materials were organized into different modules, each one composed of several lessons. Each lesson included 9 to 17 video-lectures and assessment activities. Table 1 shows the number of enrolled learners, passing rate, modules, lessons, video-lectures and assessment activities in each course. The courses followed an on-demand format in which course materials were available all at once without specific predefined deadlines. Figure 1 illustrates the structure of each course.

****

Table 1 Overview of the MOOCs in our study.

****

****

Fig. 1 MOOCs Structure. The courses are structured in modules, and each module is composed of lessons. Each lesson includes video-lectures and assessment activities. The '*' represents a video-lecture or assessment activity in each lesson.

****

2.3. Measures

Learners in the three MOOCs completed an optional questionnaire at the beginning of the course. The questionnaire included items related to demographic measures (age, gender, education) and learners' intentions in the course (to watch all lectures or only some of them). In addition, the questionnaire included the Online Learning Enrollment Intentions (OLEI) scale (Kizilcec & Schneider, 2015) translated into Spanish[2] and a measure of SRL[3]. The questions related to SRL were adapted from multiple established instruments (Littlejohn & Milligan, 2015; Barnard et al., 2008; Pintrich et al., 1991; Warr & Downing, 2000; Rigotti, Schyns, & Mohr, 2008). In total, the questionnaire included 24 statements related to six SRL strategies. Learners rated statements using a 5-point scale (coded from 0 to 4). The strategies were goal-setting strategies (4 statements), strategic planning (4), self-evaluation (3), task strategies (6), elaboration (3) and help-seeking (4). An example of a statement is, *"I read beyond the core course materials to improve my understanding."* The reliability of the final questionnaire was established in a previous study (REFERENCE REMOVED FOR BLIND REVIEW).

To categorise the students according to their SRL profiles, we used the K-Means clustering algorithm based on the similarity of students' scores in the SRL self-reported in the questionnaire. This approach grouped learners with similar SRL characteristics. Two groups of learners were identified: students with a low SRL profile (cluster 0) and students with a high SRL profile (cluster 1). The centroid obtained for cluster 0 was 2.634 and for cluster 1 it was 3.468. Similar classification techniques were employed in prior studies, which also classified learners into high and low SRL profiles according to the scores obtained in a self-reported questionnaire (Littlejohn et al., 2016; Valle et al., 2008; Romero et al., 2016).

2.4. Procedure

We used the Process Mining PM^2 method (van Eck, Lu, Leemans, & van Der Aalst, 2015), which is a simpler and more flexible adaptation of other PM methods such as the L*Life-cycle model (van Der Aalst, 2011). The PM^2 method is structured into four stages (Figure 2): (1) extraction, (2) event log generation, (3) model discovery and (4) model analysis. This method was selected because it is the one used in disciplines such as healthcare and business to understand users' interactive workflows within a particular system (Arias-Chaves & Rojas-Cordoba, 2014; Rojas, Munoz-Gama, Sepúlveda, & Capurro, 2016). It is also suitable for the analysis of both structured and unstructured processes (van Eck, Lu, Leemans & van Der Aalst, 2015).

****

Fig. 2 Stages for the generation of the process model using the PM^2 methodology. Figure adapted from van Eck et al. (2015).

****

2.4.1 Extraction Stage. In this stage, we extracted the trace data from Coursera's database in order to study the interaction sequences of learners in the MOOC. Coursera is a large platform that keeps track of almost all details of student interactions. This raw data is organized into three categories: general data, forums and personal data. It comprises 86 tables of information. For the purpose of this study, we have limited our analysis by selecting only tables (13) that contain relevant information about students' behaviour. The datasets extracted include course information, course content, course progress, assessments, course grades and learner demographics (based on user surveys).

2.4.2 Event Log Generation Stage. In this stage, we defined the event log we used in the PM algorithm. This event log is like the file collecting the information on the learners' interactions within the MOOC as well as their SRL profiles, as well as other information necessary to perform the analysis. The first step for generating the event log

[2] Spanish translation of the OLEI scale is available at http://dx.doi.org/10.6084/m9.figshare.1585144

[3] The original versions of the SRL measure questionnaire in Spanish and English are available at http://dx.doi.org/10.6084/m9.figshare.1581491

file was to define different concepts to refer to the trace data registered in the Coursera databases. Specifically, we defined the concepts of ***interaction*** and ***session*** as follows:

- An ***interaction*** is an action recorded in the Coursera trace data that registers the interaction of a learner with a MOOC object. We defined six types of interactions depending on the objects that learners interact with: start a video-lecture, complete a video-lecture, review a video-lecture already completed, try an assessment, pass an assessment, and review an assessment already passed. In addition to these interactions, we also included a label to identify the first and last interaction of the learner with the course as *begin session* and *end session*, respectively. All interactions of the learners with the MOOC content extracted from the events log are listed in Table 2.
- A ***session*** is a period of time in which the Coursera trace data registers continuous activity of a learner within the course, with intervals of inactivity no greater than 45 minutes. This definition of session was adopted from the prior works by Kovanović et al. (2015) and Liu et al. (2015).

****
Table 2 Definitions of interaction with course materials to characterize consecutive learner behaviour
****

In addition to the interactions, the event log file included the learners' SRL profiles that we extracted from the cluster indicating whether they were High or Low self-regulated learners. Finally, the event log also included whether the learner completed the course or not: a) True (finished the course), or b) False (did not finish the course). All this information is included in the event log for each session and learner. Therefore, the result of this stage is a log of events documenting the learners' interactions with the course content within a session, their self-regulated profiles, completion of the course, and other complementary data to identify the session ID, the event ID and the timestamp in which each registered event was produced. Table 3 shows an example of the event log generated.

****
Table 3 Example of the event log generated for the process analysis.
****

2.4.3. Discovery of the model. We processed the event log with a discovery algorithm to obtain a process model representing the behaviour of the learners within the MOOC. In the PM literature there is a wide range of discovery algorithms that can be used to identify interaction patterns (van Der Aalst, 2016). Given our situation, we selected the Disco algorithm (Günther and Rozinat, 2012) and Celonis algorithm and their implementations in the Disco[4] and Celonis[5] commercial tools. With some differences, both algorithms are based on the Fuzzy algorithm concept (Günther & van Der Aalst, 2007) combined with some characteristics from the Heuristic algorithm family (van Der Aalst, 2011). Both algorithms were specially designed to handle complex processes, such as learner interactions in a MOOC, and they result in process-map models that can be operated and understood by domain experts with no previous experience in PM (Günther and Rozinat, 2012). Finally, both commercial tools integrate a set of metrics and filtering options to adapt the event log to the specific questions and to analyse the process interactively. We used Disco and Celonis to generate initial process models for analysis.

2.4.4. Model analysis. Once the process model was generated, we analysed and identified learners' most frequent ***interaction sequences***. An ***interaction sequence*** is defined as a set of concatenated interactions (from one interaction to another) of the same learner within a session. That is, the path that a learner follows through the MOOC content within a session.

As a result of applying the algorithms, we obtained a *spaghetti process* model (Figure 3). The *spaghetti process model* is a term used in the PM field to refer to a model with so many arcs and crossings that it is difficult to understand or observe patterns. This process model is composed of a start-point and an end-point represented with a

[4] Disco Tool: http://www.celonis.com/en/product/

[5] Celonis Tool: https://fluxicon.com/disco/

white hexagon with a play image and a stop image inside, respectively. The interactions in Table 2 are represented with a coloured filled hexagon. The arcs and arrows connect two or more interactions into what we call *interaction sequences* that were repeated by different learners. For example, an interaction sequence would be from *Begin session* to *(→) Video-lecture-begin* to *(→) End session*, which indicates that a learner began a session, then watched a video-lecture and then ended a session; or from *Begin session* to *(→) Video-lecture-begin* to *(→) Assessment-try* to *(→) End session*, which indicates that a learner began a session, then began a video-lecture, then attempted an assessment and then ended a session. Figure 4 shows a subset of interaction sequences extracted from the main process model to provide a better explanation about its semantics. The process model also contains numbers next to each hexagon. These numbers indicate the number of times the interaction indicated in the hexagon was repeated across all sessions in the dataset. For example, Figure 4 shows that the event log contains 13,714 *Begin session* interactions; that is, there were 13,714 sessions registered in the dataset. The numbers over the arcs with arrows indicate the number of interaction sequences from the two interconnected interactions that have been identified within a session, and the arrows indicate the direction. Figure 4 shows that the interaction sequence from *Begin-session* to *(→) Video-lecture-begin* was performed 9,162 times. This means that from the 13,714 sessions that that were initiated, only 9,162 interaction sequences were performed toward *Video-lecture-begin*.

****
Fig. 3 *Spaghetti process model* containing all interaction sequences of 3 MOOCs by sessions.
****

****
Fig. 4 Representation of interaction sequences extracted from the full process model.
****

Once the process model was generated, we applied filters to the event log in order to obtain more specific process models and extract information to answer the three research questions:

- **RQ1. What are the most frequent interaction sequences of learner behaviour in a MOOC?** To answer this question, we analysed the process models in the model analysis stage to identify the most frequent interaction sequence patterns. First, we analysed the models, considering all the data from the three courses. Second, we analysed the data from each course separately.
- **RQ2. Is there any difference in the interaction sequences of learner behaviour between those who complete a course and those who do not?** To answer this question, we ran the same analysis as in RQ1, but filtered completing and non-completing learners for comparison. Also, we analysed the time spent on each interaction sequence pattern by completers and non-completers.
- **RQ3. Do interaction sequences of learner behaviour differ between learners with a high-SRL versus a low-SRL profile?** To answer this question, we ran the same analysis as in RQ1 but filtered high-SRL and low-SRL profile learners for comparison. Also, we analysed the time spent on each interaction sequence pattern by high-SRL and low-SRL learners.

3. Results

This section presents the analysis of the process models generated with the event log and the results obtained. We have organised the results according to the three research questions addressed. For each research question, we have also included an appendix consisting of a table with a set of findings, the evidence supporting the result and supporting data.

3.1. What are the most frequent interaction sequences of learner behaviour in a MOOC? *(RQ1)*

We began by analysing the process model (Figure 3) in the model analysis stage to identify the most frequent interaction sequence patterns. This process model allowed us to observe the learner behaviour as a result of the interaction with the MOOC content in a session. We found seven distinct interaction sequence patterns extracted by PM:

(1) *Only Video-lecture*: interaction sequence pattern dedicated only to watching video-lectures, in which the most common interaction sequences are *Begin session* to *video-lecture-begin* or *video-lecture-complete* or *video-lecture-review* and combinations of them before *End session* (Figure 5).

****

Fig. 5 Only Video-lecture sessions

****

(2) *Only Assessment:* interaction sequence pattern dedicated to working only with assessments in which the most common interaction sequences are *Begin session* to *assessment-try* or *assessment-pass* or *assessment-review* and combinations of them before *End session* (Figure 8 – included in the Appendix).

(3) *Assessment-try to Video-lecture*: interaction sequence pattern where the most common interaction sequences observed are (a) *Begin session* to *Assessment-try* (with the intention of trying to solve an assessment) then to *Video-lecture-begin* (looking for information in a new video-lecture) then to *Assessment-try* and *End session*, (b) *Begin session* to *Assessment-try* then to *Video-lecture-complete* (consuming the video-lecture information) then to *Assessment-try* and *End session*, and (c) *Begin session* to *Assessment-try* then to *Video-lecture-review* (looking for specific information) then to *Assessment-try* and *End session* (Figure 6 – included in the Appendix).

(4) *Video-lecture to Assessment-pass*: interaction sequence pattern where the most common interaction sequences observed are (a) *Begin session* to *Video-lecture-begin* then to *Assessment-pass* and then *End session*, (b) *Begin session* to *Video-lecture-complete* then to *Assessment-pass* and then *End session*, (c) *Begin session* to *Video-lecture-review* then to *Assessment-pass* and then *End session*, and (d) *Begin session* to *Video-lecture-begin* then to *Assessment-try* then to *Assessment-pass* and then *End session* (Figure 11 – included in the Appendix).

(5) *Video-lecture-complete to Assessment-try*: interaction sequence pattern where the most common interaction sequences observed are (a) *Begin session* to *Video-lecture-complete* then to *Assessment-try* (without achieving it and with no more attempts to complete it) and then *End session* (Figure 9 – included in the Appendix).

(6) *Explore:* interaction sequence pattern composed of an *assessment-try* and a *video-lecture-begin*, where learners only superficially inspect the contents without any intention to complete them (Figure 7 – included in the Appendix).

(7) *Composite:* interaction sequence pattern where we observed the combination of the interaction sequences mentioned before. The most common interaction sequences observed are (a) *Begin session* to various *Video-lecture-begin*s then to *Assessment-try* and then *End session* (Figure 10 – included in the Appendix).

The analysis was performed considering all the data from the three courses (Table 4) and considering the data from each course separately (Appendix – Table 11). As a result, we could confirm that the structure of the MOOC (based on its content: video-lectures and assessments) was reflected in the amount of interaction sequence patterns per session that learners performed (Appendix – Table 11).

The four most common patterns of interaction sequences among MOOC learners (93,26% of the sessions registered) are as follows, in order of frequency. (1) *Only Video-lecture* (45.25% of the sessions follow this type of pattern). The most common interaction sequence in this type of interaction pattern is *Begin session*, then *Video-lecture-begin*, then *End session* without completing the video-lecture (Appendix – Table 12 – Finding F1). **(2) *Assessment try → Video-lecture*:** 21.58% of the sessions follow this type of pattern, with the most common interaction sequence of this interaction pattern being a loop between *Begin session → Assessment-try → Video-lecture-begin → Assessment-try → Video-lecture-complete → Assessment-try → End session* (Appendix – Table 12 – Finding F2). **(3) *Explore*:** 15.67% of the sessions follow this type of pattern, in which the most common behaviour of the learners is to follow a disorganized interaction sequence in which they go from one type of content (assessments or video-lectures) to another without completing them (Appendix – Table 12 – Finding F3). **(4) *Only Assessment*:** 10.76% of the sessions follow this type of pattern, in which the most common interaction sequence is *Begin session → Assessment-try → End-session* without completing the assessment (Appendix – Table 12 – Finding F4). Finally, ***Video-lecture → Assessment-pass*** (1.10%) and ***Composite*** interaction sequence patterns are the least common (1.10% and 2.32% of the sessions, respectively) (Appendix – Table 12 – Finding F5). These patterns help us to understand how learners behave in a session, whether they complete the course or not. So, this is the starting point to study how learners with distinct profiles perform distinct interaction sequence patterns in a MOOC. In the

next section, we will analyse how distinct types of learners (categorised by course completion and SRL profile) perform these interaction patterns that provide insights about SRL strategies used throughout the course.

****

Table 4 Proportions of the interaction sequence patterns based on the number of sessions performed by learners in 3 MOOCs and derived from the MOOC process models.

****

3.2. Is there any difference in the interaction sequences of learner behaviour between those who complete a course and those who do not?

After having identified the most common interaction sequence patterns among MOOC learners in a session, we analysed how these patterns vary according to whether or not the group of learners complete the course. Specifically, we looked for differences in interaction sequence patterns that completers perform, which should help reveal how their behaviour impacts their learning and how it relates with SRL strategies. We analysed interaction sequence patterns per session and its related time. **First, we found that completers perform a higher number of assessment sessions than non-completers.** Completers' sessions mainly consist of: (a) taking one assessment after another (called *Only Assessment*) or (b) taking an assessment and then watching a video-lecture (called *Assessment try → Video-lecture*) or (c) watching video-lectures and taking an assessment without completing either (called *Explore*) or (d) combining several of the interaction sequences mentioned before (called *Composite*). By contrast, non-completers' sessions consist of watching one video-lecture after another (called *Only Video-lecture*). We found statistical differences between the percentage of sessions of each type performed by these two types of learners (Table 5). In Appendix – Table 13 – Finding F6 and Finding F7, we detailed other interaction sequence loops that characterize the behaviour of these two types of learners.

****

Table 5 Proportions of the interaction sequence patterns based on the number of sessions performed in 3 MOOCs derived from the process models for Completers and Non-Completers

Note 1: * Reject Ho: p1 = p2 and accept Ha: p1 < p2 / Note 2: ** Reject Ho: p1 = p2 and accept Ha: p1 > p2

****

Second, we found statistically significant differences between the duration of the interaction sequence patterns of completers and non-completers. While completers spend more time in sessions with assessments, non-completers invest their time in sessions focused on watching video-lectures (Table 6). Completers dedicate more time to interaction sequence patterns with a duration less than 5 minutes that consist of: (1) taking an assessment and then watching a video-lecture (called *Assessment try → Video-lecture*), (2) exploring video-lectures and assessments (called *Explore*), and (3) carrying out a complex combination of different interaction sequences (called *Composite*). They also dedicate more time to performing interaction sequence patterns (less than 5 to 10 minutes) that consist of (4) taking one assessment after another (called *Only Assessment*). Inversely, non-completers invest more time in interaction sequence patterns that consist of: (5) watching one video-lecture after another (called *Only Video-lecture*) in periods less than 5 minutes, between 5-10 minutes and over 15 minutes. The time invested in the different interaction sequence patterns is a measure to understand which actions learners put more effort into. Also, this information can help us relate interaction patterns with SRL strategies like effort regulation or elaboration. More detailed findings related to the time invested in each interaction sequence pattern are provided in Appendix – Table 13 – Findings 8 and 9.

****

Table 6 Interaction sequence patterns' duration for Completers and Non-Completers distributed per pattern.

* Statistically significant difference between proportions

** C = Completer / NC = Non-Completer

*** T1 = $t \leq 5$ min / T2 = 5 min $< t \leq 10$ min / T3 = 10 min $< t \leq 15$ min / T4 = $t >$ 15 min

****

3.3. Do interaction sequences of learner behaviour differ between learners with a high-SRL versus low-SRL profile?

We studied which interaction sequence patterns characterize those learners who self-reported high- and low-SRL profiles in the SRL questionnaire. We found differences in how SRL learners with high- and low-SRL profiles behave. **First, high-SRL learners perform a significantly higher number of sessions that include completing a video-lecture and then taking an assessment** (called *Video-lecture complete → Assessment try*) **and carrying out a complex combination of different interaction sequences** (called *Composite*) **than their counterparts.** We did not find any other statistically significant differences in other interaction sequence patterns between high- and low-SRL learners (Table 7; Appendix – Table 14 – Finding 10).

****

Table 7 Proportions of the interaction sequence patterns based on the number of sessions performed in 3 MOOCs derived from the process models for High and Low Self-Regulated Learners

Note: * Accept Ha: p1>p2 and reject Ho: p1=p2

****

Second, we found that high-SRL learners who complete the course behave differently than those learners with low-SRL profiles who also complete the course. High-SRL completers perform more Video-lectures before passing an assessment. This behaviour suggests that high-SRL completers make an effort to understand the content of the course. On the other hand, **low-SRL completers perform more assessments than video-lectures**. This suggests that these learners are more focused on passing the assessments than understanding the course content (Table 8; Appendix – Table 14 – Finding F11).

****

Table 8 Comparison of proportions of interaction sequence patterns performed by high-SRL and low-SRL completers

Note: * Accept Ha: p1>p2 and reject Ho: p1=p2 / Note 2: ** Reject Ho: p1 = p2 and accept Ha: p1 < p2

****

Third, we found statistically significant differences between the duration of the interaction sequence patterns of high-SRL completers and low-SRL completers (Table 9). High-SRL completers dedicate more time to interaction sequence patterns with a duration of less than 10 minutes that consist of: **(1) performing *Only Video-lecture* interaction sequence patterns** (where the most common interaction sequences are *Begin session* to *video-lecture-begin* or *video-lecture-complete* or *video-lecture-review* and combinations of them before *End session*) and **(2) performing complex combinations of different interaction sequences** with a duration of more than 10 minutes (where the most common interaction sequences are *Begin session* to various *Video-lecture-begin*s then to *Assessment-try* and then *End session* – called *Composite*). Low-SRL completers, on the other hand, dedicate more time to interaction sequence patterns with a duration of less than 5 minutes that consist of: **(4) taking an assessment and then watching a video-lecture**, with the most common interaction sequences being (a) *Begin session* to *Assessment-try* (with the intention of trying to solve an assessment) then to *Video-lecture-begin* (looking for information in a new video-lecture) then to *Assessment-try* and *End session*; (b) *Begin session* to *Assessment-try* then to *Video-lecture-complete* (consuming the video-lecture information) then to *Assessment-try* and *End session*; and (c) *Begin session* to *Assessment-try* then to *Video-lecture-review* (looking for specific information) then to *Assessment-try* and *End session* (called *Assessment try → Video-lecture*; Table 9; Appendix – Table 14 – Finding F12).

****

Table 9 Interaction sequence patterns' duration for High-SRL Completers (H) and Low-SRL Completers (L) distributed per pattern.

* Statistically significant difference between proportions

** H- High SRL Completer / L – Low SRL Completer

*** T1 = t ≤ 5 min / T2 = 5 min < t ≤ 10 min / T3 = 10 min < t ≤ 15 min / T4 = t > 15 min

****

4.- Discussion

This is an exploratory study in which PM techniques were applied to investigate the relationship between learners' self-reported SRL profiles and their interaction sequences captured in their interactions with the MOOC

(trace data). The incorporation of both data sources (self-reported and data traces) provides a proper measure of learners' interactions on the MOOC platform. This section presents an attempt to relate theoretical-based patterns of SRL strategies (4.1) with observed behaviour and to discuss the theoretical, practical and methodological implications (4.2).

4.1. Relating theoretical-based patterns of SRL strategies with observed behaviour

We have identified 7 interaction patterns which were defined by the most frequent interaction sequences observed from the trace data. Table 10 summarizes the relationship between these observed patterns and SRL theory.

****

Table 10 Theory-based patterns from observed behaviour related to SRL strategies

****

It is our understanding that this is the first study that makes an effort to relate behavioural patterns observed from trace data with SRL strategies in a MOOC environment. In Table 10, we have identified 7 interaction patterns. Two of these interaction patterns (*Only Video-lecture* and *Only Assessment*) are composed of either interactions with video-lectures or with assessments. The other five patterns (*Assessment try→Video-lecture*, *Video-lecture→Apass, Video-lecture-complete→Assessment try, Video-lecture-complete→Assessment try, Explore and Composite*) consist of combinations of interactions including video-lectures and assessments.

The interaction pattern *Only Video-lecture* could be related with three SRL strategies: (1) **Study SRL strategy** (Garavalia & Gredler, 2002), (2) **Rehearsal SRL strategy** (Broadbent, 2017) and (3) **Repeating SRL strategy** (Sonnenberg & Bannert, 2015). These SRL strategies are of the cognitive type. High-SRL Learners that complete the course spend periods of time between 1 and 5 minutes in this sequence pattern, which implies a kind of learning involving recall of the information rather than an effort to achieve a deep understanding of the content (Broadbent, 2017). This pattern could be complemented with information provided from an external resource (capturing trace data outside the platform), which will give us insights into whether learners take notes, draw or outline a concept map, trying to understand or better process the content. As Veletsianos et al. (2016) state, "*automatically collected data by learning platforms does not necessarily offer a comprehensive and complete representation of learners' behaviour*." This could lead to us to unveil an Organisation SRL strategy.

The interaction pattern *Only Assessment* could be related with two SRL strategies: (1) **Elaboration SRL strategy** (Weinstein et al., 2011) and (2) **Evaluation SRL strategy** (Sonnenberg & Bannert, 2015). This interaction pattern is most frequent among low-SRL learners that complete the course. Elaboration SRL strategy is of the cognitive type. The interaction pattern found in this study that relates with this SRL strategy needs to be complemented with more information about what actions learners perform in order to connect their prior knowledge with the new information. This will give us insight into their intentions in relation with how they process the information in a non-superficial way. Also, complementing this information with the time invested in the interaction with the content could led us to a better understanding of the learners' engagement with the course.

The interaction pattern *Assessment try→Video-lecture* could be related with **Help-seeking SRL strategy** (Karabenick & Dembo, 2011; Corrin, de Barba, & Bakharia, 2017). This pattern is the most common for lowSRL learners that complete the course. Help-seeking strategy aids learners when they look for help. Help seeking in online environments is generally related to looking for human help through forums, chats or other online communication mechanisms (Broadbent & Poon, 2015). But help could also come from other internal (e.g. video-lectures, forums, assessments) or external (digital or physical material outside the platform) resources. So, in order to gain a better understanding of this strategy, we need to collect invisible information that is missing in current MOOC platforms. This information could drive us to understand how the use of internal or external resources impacts learners' behaviour.

The interaction pattern *Video-lecture→Apass* could be related with **Reviewing record SRL strategy** (Zimmerman & Pons, 1986). This pattern is the most common for high-SRL learners that complete the course. This pattern is what MOOC teachers and instructional designers usually expect; students need to pass a video-lecture and then complete and pass an assessment. It could also be associated with Organization SRL strategy.

The interaction pattern *Video-lecture-complete→Assessment* try could be related with **Self-evaluation SRL strategy** (Zimmerman & Pons, 1986). This pattern is performed by High SRL learners. This SRL strategy is of the metacognitive type. The use of this SRL strategy implies that learners are proving themselves, checking their progress on the course. With the appropriate feedback, it would be possible to develop a mechanism of self-monitoring that could regulate the way in which learners arrange their learning process.

The interaction pattern *Explore* could be related with **Task exploration SRL strategy** (Van Der Linden, Sonnentag, Fresen, & van Dyck, 2010). This pattern is mainly performed by low-SRL learners that complete the course. This seems to be a strategic behaviour, consisting of jumping between video-lectures and assessments without completing them to investigate how the topics and the materials are organised.

The interaction pattern *Composite* could be related with **Effort Regulation SRL strategy** (Carson, 2011; Cho & Shen, 2013; Puzziferro, 2008). This pattern is commonly performed by high-SRL learners that complete the course. These learners tend to work for more than 10 minutes in this kind of interaction pattern, combining different interaction sequences over time and making an effort to stay engaged (focus) with the course contents. This SRL strategy is of the resource management type.

As we can see from the previous analysis, high-SRL learners that complete the course tend to follow rehearsal/repeating/study; reviewing record (organization); self-evaluation and effort-regulation SRL strategies. These learners usually follow the sequential structure proposed in the MOOC instructional design, performing more organized sessions that aim for a deep understanding of the MOOC content. Also, high-SRL completers go back and forth over the course content to review video-lectures before and after completing an assessment, behaviour that aligns with what has been called in the literature a *retrieval* SRL strategy (Johnson & Mayer, 2009; Roediger & Butler, 2011; Davis et al., 2016). Conversely, low-SRL learners that complete the course tend to use a kind of evaluation/elaboration; help-seeking and task-exploration SRL strategies. These learners are more strategic than their counterparts, since they look for specific information that will help them pass the course assessments. Finally, our results revealed that high-SRL completers also worked on the course content more intensively in terms of repetition (more interaction sequences per session) compared with low-SRL completers.

4.2. Theoretical, practical and methodological implications.

Regarding the theoretical implications, the diversity in theoretical SRL definitions and models in the last 30 years, which have attempted to describe how SRL is developed, has produced a lack of clarity in regard to SRL strategies terminology and definitions in the literature (Winters, Greene, & Costich, 2008). For example, the *Assessment try→Video-lecture* review behaviour could be associated with two kinds of SRL strategies: repeating and rehearsal. It will depend upon the theoretical SRL model adopted (e.g. the socio-cognitive model by Zimmerman and the SRL framework by Pintrich or the information process model by Winne & Hadwin). This show us that it is desirable to develop a new integrated model that includes and considers the different aspects from the existent SRL models, but is specific to the MOOC model. This new model could help developers in the design of new learning systems (such as MOOC platforms) capable of supporting the processes of SRL developed by the new integrated model.

Regarding the practical implications, the actual MOOC platforms provide trace data as a result of the learner interaction with the course content. These platforms register a great quantity of trace data, which was filtered and processed to define events that are relevant to this study's research questions. The trace data provided is not clear at all, and a great effort is needed in order to understand it (e.g. extraction and cleaning methods are required before can use the trace data). MOOC platforms should provide enriched semantic data that allow researchers to extract more understandable information about the types of interactions that learners perform on the MOOC platform, and it should be scalable for further analysis with regard to how to support SRL. Here, the level of granularity specified is an important issue for analysing SRL strategies. For example, micro-level data could improve the analysis of the development of coding schemes when learners process the content delivered in a MOOC, and macro-level data could provide insights about the process of self-regulation and the phases with which it is related. On the other hand, these levels of granularity should provide more information about SRL processes, where SRL strategies can be observed directly (rather than indirectly). A detailed taxonomy of SRL processes could support researchers in studying the role of each SRL strategy under different learning conditions (Azevedo, 2009). Thus, it is possible to build up taxonomies of different types that help us to improve our understanding of SRL from trace data. Also, common units of measure (e.g. defining a session as a period of time in which a learner develops a self-regulated process, or considering the time frame as a week or the entire course duration) are needed in order to compare the SRL processes between MOOC platforms. What's more, there should be common log files to make this comparison easier.

Regarding the methodological implications, common processes defined around SRL let researchers study it through analytical tools. These tools should complement the data obtained from self-reporting measures. This

approach is necessary and complementary; in this way it is possible to determine if what learners have self-reported is coherent and correlates with actual behaviour on the MOOC platform. Also, this information could be complemented with eye-tracking data, a seamless learning plug-in that extends the data collection to include actions that learners perform outside the platform. Both sources of data could be correlated with learning outcomes and used to quantify effective applications of SRL strategies. Harnessing learners' detailed behavioural records, unlike questionnaire data, can provide an objective longitudinal account of learning and enable real-time support and feedback. It can also help in future implementations to build tools that promote SRL in MOOCs. As Kizilcec & Brooks (2017) state, *"diverse big data and experimentation provide evidence on 'what works for whom' that can extend theories to account for individual differences and support efforts to effectively target materials and support structures in online learning environments."* In conclusion, MOOCs need to move from being content-oriented toward becoming process-oriented platforms that can better support SRL of learners.

5.- Limitations and conclusions

In this study, we have addressed three research questions to identify and characterise learners' behaviour in a MOOC focusing on SRL. We identified the following interaction sequence patterns *(RQ1)* as the most frequently repeated by learners in a MOOC: (1) watching one video-lecture after another; (2) taking one assessment after another; (3) taking an assessment and then watching a video-lecture; (4) watching a video-lecture and then passing an assessment; (5) completing a video-lecture and then taking an assessment; (6) watching video-lectures and taking an assessment without completing either; and (7) carrying out a complex combination of the different interaction sequences mentioned before. Based on these interaction patterns, we found that completers perform a higher number of sessions with assessments compared with non-completers *(RQ2)*. Also, we found statistically significant differences between the duration of the interaction sequence patterns of the completers and non-completers. In regard to the types of SRL learners *(RQ3)*, high-SRL learners perform a significantly higher number of sessions that include completing a video-lecture and then taking an assessment and carrying out a complex combination of different interaction sequences than their counterparts. Also, high-SRL learners who complete the course differ in their behaviour compared with those learners with a low-SRL profile who also complete the course. High-SRL completers perform more video-lectures before passing an assessment. This behaviour suggests that high-SRL completers make an effort to understand the content of the course. Conversely, low-SRL completers perform more assessments than video-lectures. This behaviour suggests that low-SRL completers are more strategic.

In this paper we have proposed a first attempt to relate actual observed behaviour with theory. This study is, as far as we know, the first to combine an aptitude approach with a process approach to study SRL in MOOCs and propose a theory-based pattern from observed behaviour and related to SRL strategies. To achieve this, our proposal consists of using self-reported surveys to determine learners' SRL profiles and PM techniques to analyse their interactions with the course content, extracting theory-based patterns from observed behaviour. This proposal benefits from the intersection of SRL and PM to understand how learners self-regulate in an authentic environment (Roll & Winne, 2015). This complex analysis allowed us to advance the understanding of SRL in MOOCs and to support some of the results that other researchers have observed in prior studies.

The generalisation of these results is subject to the limitations of the methodology employed in the study. Considering other data and defining the events on another level of granularity or in a distinct way could lead to different results. As in any study in which PM techniques are used, the results are directly related to the data that is taken as a base for the study and the analysis carried out for its interpretation (Bose et al., 2013). In this article, we have worked solely with one specific selection of data from Coursera. In a future project, we plan to perform the same analysis on other platforms to understand the extent to which SRL strategies are conditions of the technological environment. Recent studies point to this having an effect on learners' behaviour, but it has never been analysed on a process level to confirm it (Conole, 2015).

However, despite the inherent limitations of the methodology employed, this article contributes a new perspective and further understanding of the study of SRL in MOOCs. This study is the first to attempt to transform the information collected by a MOOC platform into event logs that register learners' interactions with the course content as events that are related to SRL strategies based on process models. Although this had been previously proposed in online studies, it has never been done in MOOCs, where the amount of data and the variety of learners is higher and more heterogeneous than in other online environments. Secondly, this is the first study that strives to transform the data from information facilitated by a platform that has not been intervened with. Until now, most research on SRL and processes carried out in online environments has been performed on platforms that were either

manipulated or adapted to support SRL, by adding functionalities that were directly associated with a self-regulated strategy (Sonnenberg & Bannert, 2015; Beheshitha, Gašević & Hatala, 2015). This study, however, is based on information from a platform that has not been manipulated for this aim. Lastly, both the definition of events and applied methodology were presented with the necessary accuracy to be reproducible. The aim of this study is to serve as a reference for other researchers who would like to analyse their courses, combining an aptitude and PM approach to further the understanding of how students learn in MOOCs. This exploratory study opens the discussion to the theoretical, practical and methodological implications for developing these kinds of studies.

References

Arias Chaves, M., & Rojas Cordoba, E. (2014). Deciphering event logs in SharePoint Server: A methodology based on process mining. In *Computing Conference (CLEI), 2014 XL Latin American* (pp. 1–12).

Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. Metacognition and Learning, 4(1), 87-95.

Bannert, M. (2009). Promoting self-regulated learning through prompts. *Zeitschrift Für Pädagogische Psychologie*, *23*(2), 139–145.

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.

Barnard, L., Paton, V., & Lan, W. (2008). Online self-regulatory learning behaviours as a mediator in the relationship between online course perceptions with achievement. The International Review of Research in Open and Distributed Learning, 9(2).

Beheshitha, S. S., Gašević, D., & Hatala, M. (2015). A process mining approach to linking the study of aptitude and event facets of self-regulated learning. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (pp. 265–269).

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, *7*(2), 161–186.

Boekaerts, M. (1999). Self-regulated learning: where we are today. *International Journal of Educational Research*, *31*, 445–457.

Borkowski, J. G. (1996). Metacognition: ¿Theory or chapter heading? *Learning and Individual Differences*, *8*(4), 391–402.

Bose, R. P., Mans, R. S., & van Der Aalst, W. M. P. (2013). Wanna improve process mining results? In Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on (pp. 127–134).

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G., Ho, A. D., & Seaton, D. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, *8*(March 2012), 13–25. http://www.mendeley.com/catalog/studying-learning-worldwide-classroom-research-edxs-first-mooc/. Accessed 13 October 2015

Broadbent, J. (2017). Comparing online and blended learner's self-regulated learning strategies and academic performance. The Internet and Higher Education, 33, 24-32.

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. The Internet and Higher Education, 27, 1-13.

Carson, A. D. (2011). Predicting student success from the LASSI for learning online (LLO). Journal of Educational Computing Research, 45(4), 399–414.

Cho, M.-H., & Shen, D. (2013). Self-regulation in online learning. Distance education, 34(3), 290–301.

Conole, G. (2015). Designing effective MOOCs. Educational Media International, 52(4), 239-252.

Cooper, S., & Sahami, M. (2013). Reflections on Stanford's MOOCs. *Communications of the ACM*, *56*(2), 28–30. article.

Corrin, L., de Barba, P. G., & Bakharia, A. (2017). Using learning analytics to explore help-seeking learner profiles in MOOCs. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference (pp. 424-428). ACM.

Davis, D., Chen, G., van der Zee, T., Hauff, C., & Houben, G.-J. (2016). Retrieval Practice and Study Planning in MOOCs: Exploring Classroom-Based Self-regulated Learning Strategies at Scale. In *European Conference on Technology Enhanced Learning* (pp. 57–71). inproceedings.

Daradoumis, T., Bassi, R., Xhafa, F., & Caballe, S. (2013). A Review on Massive E-Learning (MOOC) Design, Delivery and Assessment. *2013 Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, 208–213. doi:10.1109/3PGCIC.2013.37

Dietze, S., Siemens, G., Taibi, D., & Drachsler, H. (2016). Editorial: Datasets for Learning Analytics. Journal of Learning Analytics, 3(2), 307-311.

Eynon, R. (2013). The rise of Big Data: what does it mean for education, technology, and media research?

Garavalia, L. S., & Gredler, M. E. (2002). Prior achievement, aptitude, and use of learning strategies as predictors of college student achievement. *College Student Journal*, *36*(4), 616.

Gasevic, D., Kovanovic, V., Joksimovic, S., & Siemens, G. (2014). Where is research on massive open online courses headed? A data analysis of the MOOC Research Initiative. *The International Review of Research In Open And Distributed Learning*, *15*(5).

Günther, C. W., & Rozinat, A. (2012). Disco: Discover Your Processes. *BPM (Demos)*, *940*, 40–44. article.

Günther, C. W., & van Der Aalst, W. M. P. (2007). Fuzzy mining--adaptive process simplification based on multi-perspective metrics. In *Business Process Management* (pp. 328–343). Springer.

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. Metacognition and Learning, 2(2-3), 107–124.

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

Jansen, R.S., van Leeuwen, A., Janssen, J. (2016). Validation of the self-regulated online learning questionnaire. Journal of Computing in Higher Education, 1-22, Springer, doi:10.1007/s12528-016-9125-x

Jivet, I. (2016). *The Learning Tracker*. *A Learner Dashboard that Encourages Self-Regulation in MOOC Learners*. TU Delft. Retrieved from http://repository.tudelft.nl/

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. Journal of Educational Psychology, 101(3), 621.

Johnson, A. M., Azevedo, R., & D'Mello, S. K. (2011). The temporal and dynamic nature of self-regulatory processes during independent and externally assisted hypermedia learning. *Cognition and Instruction*, *29*(4), 471–504.

Karabenick, S. A., & Dembo, M. H. (2011). Understanding and facilitating self-regulated help seeking. New Directions for Teaching and Learning, 2011(126), 33-43.

Kizilcec, R. F. & Brooks, C. (2017- in press). Diverse Big Data and Randomized Field Experiments in Massive Open Online Courses: Opportunities for Advancing Learning Research. In G. Siemens & C. Lang (Eds.), Handbook on Learning Analytics & Educational Data Mining.

Kizilcec, R. F., & Schneider, E. (2015). Motivation as a Lens to Understand Online Learners: Toward Data-Driven Design with the OLEI Scale. *Transactions on Computer-Human Interactions (TOCHI)*, *22*(2), 24.

Kizilcec, R., Pérez-Sanagustín, M., & Maldonado, J. J. (2016). Self-Regulated Learning in Massive Open Online Courses: Individual Differences and a Study Tips Experiment. In *Learning at Scale Conference 2016*.

Kizilcec, R., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. Computers & Education, 2017.

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S., & Hatala, M. (2015). Penetrating the black box of time-on-task estimation. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 184–193). inproceedings.

Littlejohn, A., Hood, N., Milligan, C., & Mustain, P. (2016). Learning in MOOCs: Motivations and self-regulated learning in MOOCs. The Internet and Higher Education, 29, 40–48.

Littlejohn, A., & Milligan, C. (2015). Designing MOOCs for professional learners: tools and patterns to encourage self-regulated learning. eLearning Papers. eLearning Papers.

Liu, Z., He, J., Xue, Y., Huang, Z., Li, M., & Du, Z. (2015). Modeling the learning behaviors of massive open online courses. In *Big Data (Big Data), 2015 IEEE International Conference on* (pp. 2883–2885). inproceedings.

Lodge, J. M., & Corrin, L. (2017). What data and analytics can and do say about effective learning. *npj Science of Learning*, 2(1), 5.

Lodge, J. M. & Lewis, M. J. (2012). In Future Challenges, Sustainable Futures. *Proceedings ascilite Wellington 2012 (eds*. *Brown, M., Hartnett, M., & Stewart, T.)*

Mukala, P., Buijs, J., & van Der Aalst, W. M. P. (2015a). Exploring students' learning behaviour in moocs using process mining techniques. Retrieved from http://www.bmpcenter.org

Mukala, P., Buijs, J., & van Der Aalst, W. M. P. (2015b). Uncovering learning patterns in a mooc through conformance alignments. Retrieved from http://www.bmpcenter.org

Pintrich, P. R. (1991). A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich & M. Zeidner (eds), Handbook of Self-regulation. San Diego, CA: Academic Press.

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, *16*(4), 385–407.

Puzziferro, M. (2008). Online technologies self-efficacy and self-regulated learning as predictors of final grade and satisfaction in college-level online courses. The Amer. Jrnl. of Distance Education, 22(2), 72–89.

Rigotti, T., Schyns, B., & Mohr, G. (2008). A short version of the occupational self-efficacy scale: Structural and construct validity across five countries. Journal of Career Assessment, 16(2), 238-255.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. Trends in Cognitive Sciences, 15(1), 20–27.

Rojas, E., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2016). Process Mining in Healthcare: A literature review. *Journal of Biomedical Informatics*. Retrieved from http://www.sciencedirect.com/science/article/pii/S1532046416300296

Roll, I., & Winne, P. H. (2015). Understanding, evaluating, and supporting self-regulated learning using learning analytics. Journal of Learning Analytics, 2, 7–12.
Romero, C., Cerezo, R., Bogarin, A., & Sánchez-Santillán, M. (2016). Educational process mining: a tutorial and case study using moodle data sets. Data Mining and Learning Analytics: Applications in Educational Research, 1.
Roth, A., Ogrin, S., & Schmitz, B. (2015). Assessing self-regulated learning in higher education: a systematic literature review of self-report instruments. *Educational Assessment, Evaluation and Accountability*, 1–26.
Siadaty, M., Gašević, D., & Hatala, M. (2016). Measuring the impact of technological scaffolding interventions on micro-level processes of self-regulated workplace learning. *Computers in Human Behavior*, *59*, 469–482.
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.
Valle, A., Núñez, J. C., Cabanach, R. G., González-Pienda, J. A., Rodríguez, S., Rosário, P., et al. (2008). Self-regulated profiles and academic achievement. *Psicothema*, *20*(4), 724–731. article.
van Der Aalst, W. (2016). *Process mining: discovery, conformance and enhancement of business processes* (Second Edi.). book, Springer Science & Business Media.
van Der Aalst, W. (2011). *Process mining: discovery, conformance and enhancement of business processes*. Springer Science & Business Media.
van Der Aalst, W., Adriansyah, A., de Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., … others. (2011). Process mining manifesto. In Business process management workshops (pp. 169–194).
Van Der Linden, D., Sonnentag, S., Frese, M., & Van Dyck, C. (2010). Exploration strategies, performance, and error consequences when learning a complex computer task. Behaviour & Information Technology, 20(3), 189-198.
van Eck, M. L., Lu, X., Leemans, S. J. J., & van Der Aalst, W. M. P. (2015). PM^ 2: A Process Mining Project Methodology. In *Advanced Information Systems Engineering* (pp. 297–313).
Veletsianos, G., Reich, J., & Pasquini, L. A. (2016). The Life Between Big Data Log Events: Learners' Strategies to Overcome Challenges in MOOCs. AERA Open, 2(3), 2332858416657002.
Warr, P., & Downing, J. (2000). Learning strategies, learning anxiety and knowledge acquisition. British journal of Psychology, 91(3), 311-333.
Weinstein, C. E., Acee, T. W., & Jung, J. (2011). Self-regulation and learning strategies. New Directions for Teaching and Learning, 2011(126), 45–53.
Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist*, *45*(4), 267–276.
Winne, P. H. (2014). Issues in researching self-regulated learning as patterns of events. *Metacognition and Learning*, *9*(2), 229–237.
Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. *Metacognition in Educational Theory and Practice*, *93*, 27–30.
Winters, F. I., Greene, J. A., & Costich, C. M. (2008). Self-regulation of learning within computer-based learning environments: A critical analysis. Educational Psychology Review, 20(4), 429-444.
Wirth, J., & Leutner, D. (2008). Self-regulated learning as a competence: Implications of theoretical models for assessment methods. *Zeitschrift Für Psychologie/Journal of Psychology*, *216*(2), 102–110.
Zimmerman, B. J., & Pons, M. M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. American Educational Research Journal, 23(4), 614–628. article.
Zimmerman, B. J. (1998). Developing self-fulfilling cycles of academic regulation: An analysis of exemplary instructional models. In D. H. Schunk, & B. J. Zimmerman, (Eds.), Self-regulated learning: From teaching to self-reflective practice (pp. 1-19). New York: Guilford Press.
Zimmerman, B. J. (2000). Attaining Self-Regulation: a social cognitive perspective. *Handbook of Self-Regulation*, 13–39.
Zimmerman, B. J. (2015). *Self-Regulated Learning: Theories, Measures, and Outcomes*. *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier. Retrieved from http://www.sciencedirect.com/science/article/pii/B9780080970868260601

Acknowledgments

This work was supported by FONDECYT (Chile) under project initiation grant N 11150231, the MOOC-Maker Project grant N 561533-EPP-1-2015-1-ES-EPPKA2-CBHE-JP), and the Comisión Nacional de Investigación Científica – CONICYT/ DOCTORADO NACIONAL 2016/21160081, Ministry of Education, Chile, Ph.D. Student Fellowships and University of Cuenca, Ecuador.

Tables

TABLES MANUSCRIPT: Mining Theory-Based Patterns from Big Data: Identifying Self-Regulated Learning Strategies in Massive Open Online Courses

Table 1 Overview of the MOOCs in our study.

	MOOC 1 (n=497)	MOOC 2 (n=2,035)	MOOC 3 (n=926)
Enrolled	18,653	25,706	10,576
Passing Rate	1,40%	8,40%	11,40%
Modules	9	4	7
Lessons	9	17	13
Video-lectures	48	83	51
Assessments	7	16	6

Table 2 Definitions of interaction with course materials to characterize consecutive learner behaviour

Interaction	Definition
Begin session	First interaction with a MOOC object in a session.
End session	Last interaction with a MOOC object in a session.
Video-Lecture begin	Begin watching a video-lecture without completing it. The video-lecture was not previously completed.
Video-Lecture complete	Watch a video-lecture in its entirety on the first attempt.
Video-Lecture review	Go back to a video-lecture that the learner had previously watched in its entirety (not necessarily on the first attempt).
Assessment try	Unsuccessful attempt to solve an assessment.
Assessment pass	Successful attempt to solve an assessment for the first time.
Assessment review	Go back to an assessment that was previously completed successfully (not necessarily on the first attempt).

Table 3 Example of the event log generated for the process analysis.

Case ID	Time Stamp	Interaction	High SRL (profile)	Course completion	Session
c7a1821f350de427f31acc92cf40b27c8a36ea9d	1451023929	Begin session	False	False	1
c7a1821f350de427f31acc92cf40b27c8a36ea9d	1448567431	Video-Lecture.begin	False	False	1
c7a1821f350de427f31acc92cf40b27c8a36ea9d	1448567737	Video-Lecture.complete	False	False	2
c7a1821f350de427f31acc92cf40b27c8a36ea9d	1448568139	Assessment.try	False	False	2
c7a1821f350de427f31acc92cf40b27c8a36ea9d	1449103918	Video-Lecture.repeat	False	False	1
011ff41dfa7cc2cf9bb89a73fd9ac1ac74eef4d3	1449104348	Assessment.pass	True	True	1
011ff41dfa7cc2cf9bb89a73fd9ac1ac74eef4d3	1449104694	Assessment.review	True	True	2
011ff41dfa7cc2cf9bb89a73fd9ac1ac74eef4d3	1449105157	End session	True	True	1
.....					

Table 4 Proportions of the interaction sequence patterns based on the number of sessions performed by learners in 3 MOOCs and derived from the MOOC process models.

Interaction sequence patterns	ALL 3 MOOCS		
	No.**	**%**	**Learners**
Only Video-lecture	6,206	**45.25**	2,495
Atry → Video-lecture	2,96	**21.58**	1,271
Explore	2,149	**15.67**	1,195
Only Assessment	1,475	**10.76**	865
Video-lecture complete → Atry	455	**3.32**	358
Composite	318	**2.32**	258
Video-lecture → Apass	151	**1.10**	132
Total	**13,714**	**100%**	**-**

** No. refers to the number of sessions in which each interaction sequence pattern is present.

Table 5 Proportions of the interaction sequence patterns based on the number of sessions performed in 3 MOOCs derived from the process models for Completers and Non-Completers

Note 1: * Reject Ho: p1 = p2 and accept Ha: p1 < p2 / Note 2: ** Reject Ho: p1 = p2 and accept Ha: p1 > p2

Interaction sequence patterns	COMPLETER			NON-COMPLETER			Ztest	P-value
	No.	**%**	**Learners**	**No.**	**%**	**Learners**		
Only Video-lecture	1,253	**36.29**	240	4,953	**48.27**	2,255	-12.23	**0.000***
Atry → Video-lecture	922	**26.70**	228	2,038	**19.86**	1,043	8.450	**0.000****
Explore	610	**17.67**	208	1,539	**15.00**	987	3.729	**0.000****
Only Assessment	417	**12.08**	169	1,058	**10.31**	696	2.896	**0.001****
Video-lecture complete → Atry	111	**3.21**	77	344	**3.35**	281	-0.391	0.347
Composite	96	**2.78**	67	222	**2.16**	191	2.082	**0.018****
Video-lecture → Apass	44	**1.27**	38	107	**1.04**	94	1.127	0.129
Total	**3,453**	**100%**	**-**	**10,261**	**100%**	**-**	**-**	**-**

** No. refers to the number of sessions in which each interaction sequence pattern is present.

Table 6 Interaction sequence patterns' duration for Completers and Non-Completers distributed per pattern.
* Statistically significant difference between proportions
** C = Completer / NC = Non-Completer
*** T1 = t ≤ 5 min / T2 = 5 min < t ≤ 10 min / T3 = 10 min < t ≤ 15 min / T4 = t > 15 min

Interaction sequences patterns' duration	**T1*** **				**T2*** **				**T3*** **				**T4*** **			
Interaction sequences patterns	**Percentage of sessions per time**															
	C**	**NC****	**Ztest**	**P-value**	**C****	**NC****	**Ztest**	**P-value**	**C****	**NC****	**Ztest**	**P-value**	**C****	**NC****	**Ztest**	**P-value**
Only Video-lecture	49.6%	60.2%	-8.83	**0.0001***	17.1%	19.7%	-1.59	0.1124	3.7%	8.6%	-2.53	**0.0057***	0%	3.5%	-2.12	**0.017***
Atry → Video-lecture	13.1%	10.3%	3.68	**0.0001***	38.4%	37.8%	0.31	0.7602	72.9%	67.2%	1.62	0.0524	89.6%	83.1%	1.53	0.125
Explore	14.9%	11.6%	4.1	**0.0001***	29.7%	28.0%	0.92	0.1786	10.4%	12.3%	-0.79	0.4289	1.6%	0.0%	-	-
Only Assessment	18.0%	13.8%	4.83	**0.0001***	2.1%	1.3%	1.75	**0.0405***	0.4%	0.2%	0.41	0.6798	0%	0.7%	-	-
Video-lecture complete → Atry	2.6%	2.8%	-0.56	0.2869	5.6%	5.7%	-0.11	0.9097	2.2%	2.1%	0.13	0.9002	0.8%	0.7%	0.09	0.9278
Composite	1.1%	0.7%	1.92	**0.0274***	5.0%	5.5%	-0.57	0.5706	7.8%	7.5%	0.14	0.8918	6.4%	8.5%	-0.64	0.5253
Video-lecture → Apass	0.8%	0.6%	0.84	0.2001	2.1%	2.1%	0	0.9987	2.6%	2.1%	0.45	0.6535	1.6%	3.5%	-0.98	0.3269
Total	**100%**	**100%**	**-**	**-**	**100%**	**100%**	**-**	**-**	**100%**	**100%**	**-**	**-**	**100%**	**100%**	**-**	**-**

Table 7 Proportions of the interaction sequence patterns based on the number of sessions performed in 3 MOOCs derived from the process models for High and Low Self-Regulated Learners
Note: * Accept Ha: p1>p2 and reject Ho: p1=p2

Interaction sequence patterns	HIGH SRL			LOW SRL			Ztest	P-value
	No.	%	Learners	No.	%	Learners		
Only Video-lecture	3,267	**45.26**	1,272	2,939	**45.25**	1,223	0.063	0.9950
Atry → Video-lecture	1,537	**21.29**	643	1423	**21.91**	628	-0.878	0.3797
Explore	1,109	**15.36**	619	1040	**16.01**	576	-1.045	0.2957
Only Assessment	752	**10.42**	456	723	**11.13**	409	-1.348	0.1774
Video-lecture complete → Atry	267	**3.70**	207	188	**2.89**	151	2.625	**0.0043***
Composite	198	**2.74**	159	120	**1.85**	99	3.477	**0.0003***
Video-lecture → Apass	89	**1.23**	78	62	**0.95**	54	1.559	0.0595
Total	**7,219**	**100%**	**-**	**6,495**	**100%**	**-**	**-**	**-**

** No. refers to the number of sessions in which each interaction sequence pattern is present.

Table 8 Comparison of proportions of interaction sequence patterns performed by high-SRL and low-SRL completers
Note: * Accept Ha: p1>p2 and reject Ho: p1=p2 / Note 2: ** Reject Ho: p1 = p2 and accept Ha: p1 < p2

Interaction sequence patterns	HIGH-SRL Completer		LOW-SRL Completer			
	No.	%	No.	%	Ztest	P-value
Only Video-lecture	3,089	**26.66**	2,071	**23.48**	5.183	**0,0001***
Atry → Video-lecture	4,524	**39.05**	3,641	**41.28**	-3.22	**0.0012****
Explore	1,895	**16.36**	1,665	**18.88**	-4.698	**0.0001****
Only Assessment	809	**6.98**	714	**8.09**	-2.994	**0.0027****
Video-lecture complete → Atry	404	**3.49**	286	**3.24**	0.96	0,338
Composite	651	**5.62**	319	**3.62**	6.66	**0.0001***
Video-lecture → Apass	214	**1.85**	125	**1.42**	2.38	**0.0086***
Total	**11,586**	**100%**	**8,821**	**100%**	**-**	**-**

** No. refers to the number of sessions in which each interaction sequence pattern is present.

Table 9 Interaction sequence patterns' duration for High-SRL Completers (H) and Low-SRL Completers (L) distributed per pattern.
* Statistically significant difference between proportions
** H- High SRL Completer / L – Low SRL Completer
*** T1 = t ≤ 5 min / T2 = 5 min < t ≤ 10 min / T3 = 10 min < t ≤ 15 min / T4 = t > 15 min

Interaction sequence patterns' duration	**T1*** **				**T2*** **				**T3*** **				**T4*** **			
Interaction sequence patterns	**Percentage of sessions per time**															
	H**	**L****	**Ztest**	**P-value**	**H****	**L****	**Ztest**	**P-value**	**H****	**L****	**Ztest**	**P-value**	**H****	**L****	**Ztest**	**P-value**
Only Video-lecture	51.8%	47.0%	2.23	**0.013***	19.7 %	13.8%	2.25	**0.012***	4.8%	1.9%	1.24	0.216	0%	0%	-	-
Atry → Video-lecture	11.8%	14.6%	-1.97	**0.024***	36.9 %	40.5%	-1.08	0.28	69.7%	77.9%	-1.47	0.141	89%	90.4 %	-0.24	0.808
Explore	13.9%	16.0%	-1.35	0.177	29.6 %	29.8%	-0.05	0.963	7.9%	14.4%	-1.71	0.087	2.7%	0%	1.20	0.228
Only Assessment	17.2%	18.9%	-1.07	0.284	1.7%	2.8%	-1.1	0.273	0.6%	0.0%	0.8	0.4	0%	0%	-	-
Video-lecture complete → Atry	3.1%	2.0%	1.65	0.099	5.0%	6.3%	-0.86	0.39	2.4%	1.9%	0.27	0.786	0%	1.9%	1.19	0.234
Composite	1.3%	0.9%	0.99	0.321	5.6%	4.1%	0.97	0.333	10.9%	2.9%	2.39	**0.008***	6.9%	5.8%	0.24	0.807
Video-lecture → Apass	0.9%	0.6%	0.87	0.381	1.7%	2.8%	-1.1	0.273	3.6%	1.0%	1.34	0.179	1.4%	1.9%	-0.24	0.808
Total	**100%**	**100%**	**-**	**-**	**100 %**	**100%**	**-**	**-**	**100%**	**100%**	**-**	**-**	**100 %**	**100 %**	**-**	**-**

Table 10 Theory-based patterns from observed behaviour related to SRL strategies

Interaction patterns	Description	SRL Strategy
Only Video-lecture	Interaction pattern dedicated to working only with video-lectures (2 or more consecutively). The interaction sequence patterns consist of: *Begin session* to *video-lecture-begin* or *video-lecture-complete* or *video-lecture-review* and combinations of them before *End session*.	The interaction sequences referring to *video-lecture begin* and *video-lecture complete* could be related to the ***Study*** SRL strategy described by Garavalia & Gredler (2002) (e.g. "Study in a particular order"). *Video-lecture review* in isolation is related to the ***Rehearsal*** SRL strategy described by Broadbent (2017) (e.g. "Learner who listens to an online lecture repeatedly") or by Weinstein et al. (2011) (e.g. "Go over information"). This pattern could also be related to ***Repeating,*** an SRL strategy defined by Sonnenberg & Bannert (2015) as "Watching (part of) a lecture that was completed in the past."
Only Assessment	Interaction pattern dedicated to working only with assessments (2 or more consecutively). The interaction sequences patterns consist of: *Begin session* to *assessment-try* or *assessment-pass* or *assessment-review* and combinations of them before *End session*.	The interaction sequences referring to *assessment-try* and *assessment-pass* could be related with the ***Elaboration*** SRL strategy described by Weinstein et al. (2011) (e.g. "Answering possible test questions"). When assessment review occurs it could also be associated with the ***Evaluation*** SRL strategy described by Sonnenberg & Bannert (2015) (e.g. "Look up an assessment that was completed in the past").
Assessment try →Video-lecture	Interaction pattern where the learner tries an assessment and then performs a video-lecture interaction. The interaction sequence patterns consist of: (a) *Begin session* to *Assessment-try* (with the intention of trying to solve an assessment) then to *Video-lecture-begin* (looking for information in a new video-lecture) then to *Assessment-try* and *End session*. (b) *Begin session* to *Assessment-try* then to *Video-lecture-complete* (consuming the video-lecture information) then to *Assessment-try* and *End session*. (c) *Begin session* to *Assessment-try* then to *Video-lecture-review* (looking for specific information) then to *Assessment-try* and *End session*.	These interaction sequences (a), (b) and (c) could be associated with the ***Help-seeking*** SRL strategy (Karabenick & Dembo, 2011; Corrin, de Barba, & Bakharia, 2017). This help-seeking could be classified as internal if the learner looks for information inside the MOOC environment, or as external if they look for information outside the MOOC platform, using resources such as web pages, digital books, learning objects, etc.
Video-lecture→Apass	Interaction pattern where the learner passes an assessment after performing many video-lecture interactions. The interaction sequence patterns consist of: (a) *Begin session* to *Video-lecture-begin* then to *Assessment-pass* and then *End session*. (b) *Begin session* to *Video-lecture-complete* then to *Assessment-pass* and then *End session*. (c) *Begin session* to *Video-lecture-review* then to *Assessment-pass* and then *End session*. (d) *Begin session* to *Video-lecture-begin* then to *Assessment-try* then to *Assessment-pass* and then *End session*.	The interaction sequences performed in (b) correspond to those proposed in the MOOC instructional design in the MOOC platform (*Video-lecture-complete→Apass*). Interaction sequences (a), (b), (c) and (d) could be associated with the ***Reviewing record*** SRL strategy described by Zimmerman & Pons (1986) (e.g. "Learner initiated efforts to try, complete or review test, notes, or textbooks to prepare for a test").
Video-lecture-complete → Assessment try	Interaction pattern where the learner attemptsto solve an assessment after completing a video-lecture. This interaction sequence pattern consists of: *Begin session* to *Video-lecture-complete* then to *Assessment-try* (without achieving it and with no more intentions made to complete it) and then *End session*.	This interaction pattern could be associated with the ***Self-evaluation*** SRL strategy described by Zimmerman & Pons (1986) (e.g. "Student initiated evaluations of the progress of their work").

Explore	Interaction pattern performed by lurker learners, who only superficially inspect the video-lectures and assessments (*video-lecture begin* and *assessment try*) without any intention to complete them.	This interaction pattern could be associated with the ***Task exploration*** SRL strategy described by Van Der Linden, Sonnentag, Fresen, & van Dyck (2010) (e.g. "The task exploration strategies performed in order to obtain more information and plan for learning a new computer program").
Composite	Interaction pattern where more than one of the aforementioned interaction patterns are performed in combination.	When learners perform a subset of different interaction sequence patterns continuously for an extended period of time, it could be associated with the ***Effort-regulation*** SRL strategy, which has been correlated with high academic achievements in online learning environments (Carson, 2011; Cho & Shen, 2013; Puzziferro, 2008).

Figures

FIGURES MANUSCRIPT: Mining Theory-Based Patterns from Big Data: Identifying Self-Regulated Learning Strategies in Massive Open Online Courses

MOOC 1: Aula Constructivista			MOOC 2: Electrones en Accion			MOOC 3: Gestion de Organizaciones		
	Video-Lecture	Assessm.		Video-Lecture	Assessm.		Video-Lecture	Assessm.
Module 1			Module 1			Module 1		
Lesson 1	**		Lesson 1	*		Lesson 1	**	
			Lesson 2	****	*			
			Lesson 3	***	*			
			Lesson 4	**	*			
			Lesson 5	**	*			
Module 2			Module 2			Module 2		
Lesson 1	*****		Lesson 1	**********	*	Lesson 1	*****	
Lesson 2		*	Lesson 2	*********	*	Lesson 2	**	*
			Lesson 3	********	*			
			Lesson 4	********	*			
Module 3			Module 3			Module 3		
Lesson 1	******		Lesson 1	*****	*	Lesson 1	*****	
Lesson 2		*	Lesson 2	****	*	Lesson 2	**	*
			Lesson 3	****	*			
			Lesson 4	****	*			
Module 4			Module 4			Module 4		
Lesson 1	********		Lesson 1	*******	*	Lesson 1	******	
Lesson 2		*	Lesson 2	***	*	Lesson 2	**	*
			Lesson 3	*****	*			
			Lesson 4	****	*			
Module 5						Module 5		
Lesson 1	*****					Lesson 1	**********	
Lesson 2		*				Lesson 2	**	*
Module 6						Module 6		
Lesson 1	*******					Lesson 1	*****	
Lesson 2		*				Lesson 2	**	*
Module 7						Module 7		
Lesson 1	**********					Lesson 1	*****	
Lesson 2		*				Lesson 2	**	*
Module 8								
Lesson 1	*****							
Lesson 2		*						
Module 9								
Lesson 1	*							

Fig. 1 MOOCs Structure. The courses are structured in modules, and each module is composed of lessons. Each lesson includes video-lectures and assessment activities. The '*' represents a video-lecture or assessment activity in each lesson.

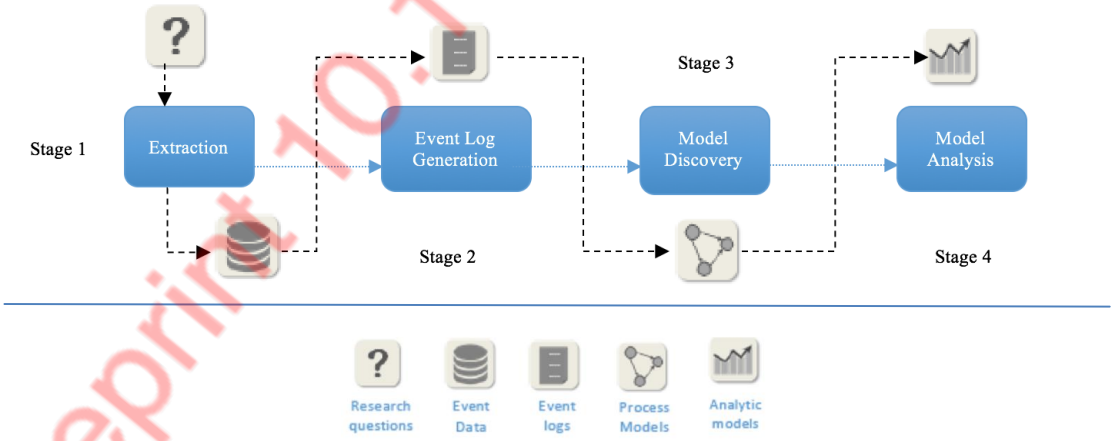

Fig. 2 Stages for the generation of the process model using the PM^2 methodology. Figure adapted from van Eck et al. (2015).

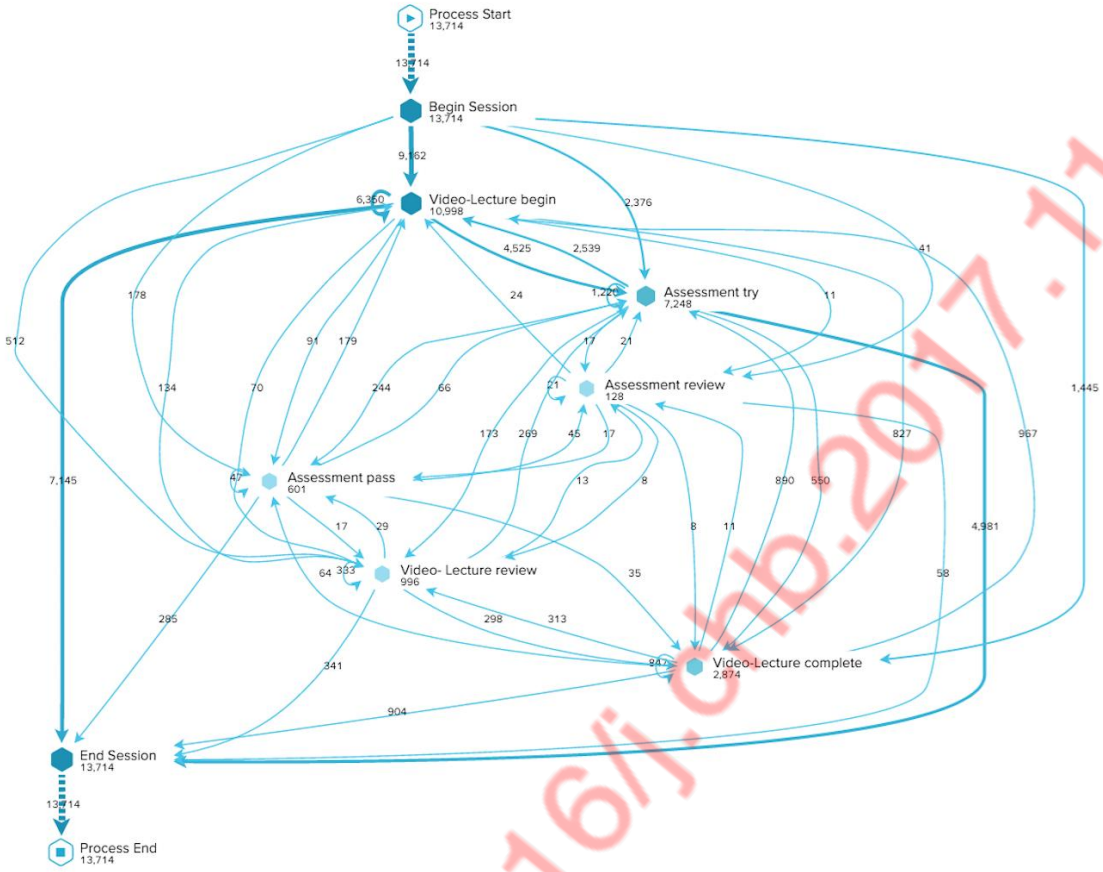

Fig. 3 *Spaghetti process model* containing all interaction sequences of 3 MOOCs by sessions.

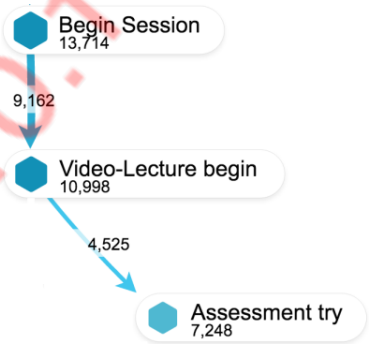

Fig. 4 Representation of interaction sequences extracted from the full process model.

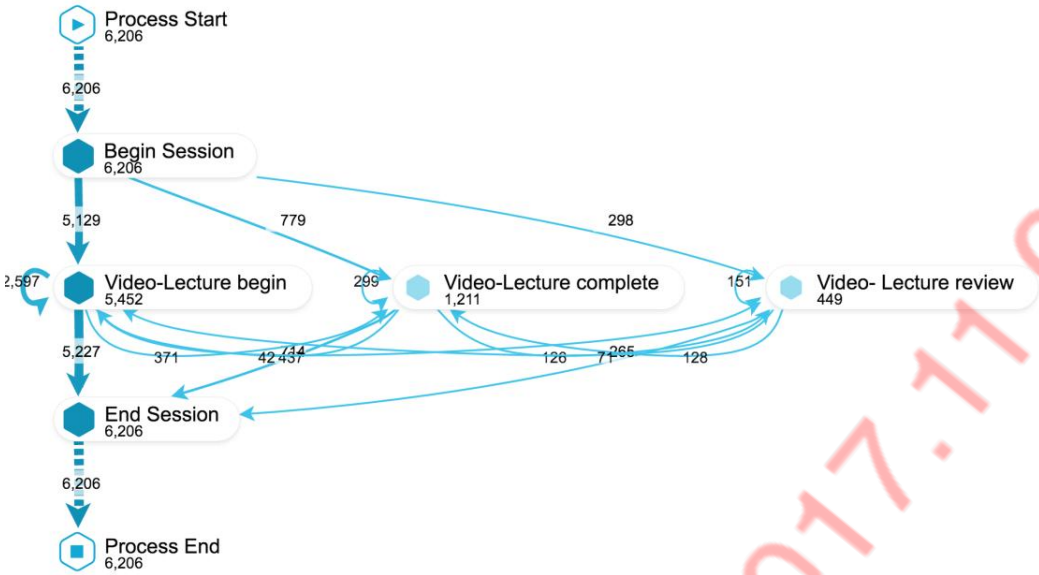

Fig. 5 Only Video-lecture sessions

Suplementary Material - Appendix

Click here to download Supplementary Material (for online Publication): Appendix.docx