Computers in Human Behavior 80 (2018) 179–196

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses

Jorge Maldonado-Mahauad [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, Avda. Vicuña Mackenna 4860, Macul, Santiago, Chile
[b] Universidad de Cuenca, Department of Computer Science, Av. 12 Abril, Cuenca, Ecuador
[c] Stanford University, Graduate School of Education, 485 Lausen Mall, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 28 March 2017
Received in revised form
9 November 2017
Accepted 10 November 2017
Available online 13 November 2017

Keywords:
Self-regulated learning
Learning strategies
Process mining
Massive open online courses

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 3458 learners across three Massive Open Online Courses. We identified six distinct interaction sequence patterns. We matched each interaction sequence pattern with one or more theory-based SRL strategies and identified three clusters of learners. First, Comprehensive Learners, who follow the sequential structure of the course materials, which sets them up for gaining a deeper understanding of the content. Second, Targeting Learners, who strategically engage with specific course content that will help them pass the assessments. Third, Sampling Learners, who exhibit more erratic and less goal-oriented behaviour, report lower SRL, and underperform relative to both Comprehensive and Targeting Learners. Challenges that arise in the process of extracting theory-based patterns from observed behaviour are discussed, including analytic issues and limitations of available trace data from learning platforms.

1. Introduction

In recent years, masses of fine-grained educational records have become available to researchers and accelerated the nascent field of learning analytics (Dietze, Siemens, Taibi, & Drachsler, 2016). Digital learning platforms collect detailed records of each learner's behaviour, performance, and other types of interaction. In particular, Massive Open Online Courses (MOOCs) are a major source of data on learner behaviour and they enable research to gain a better understanding of how individuals learn in online learning environments (Breslow et al., 2013; Cooper & Sahami, 2013; Daradoumis, Bassi, Xhafa, & Caballe, 2013).

Nevertheless, despite the large amount of data that MOOCs are collecting, this information may not be sufficient to build on educational theories and develop new ones. In particular, access to critical information about learners' behaviour and learning processes is frequently limited (Lodge & Lewis, 2012). Data-driven methods can rapidly extract patterns in what learners do throughout a course, but it remains a challenge to interpret the patterns and understand how they relate to theory. One approach to increase the interpretability of large amounts of clickstream data is to triangulate with other data sources (i.e., taking a mixed-methods approach). For example, clickstream data from MOOCs, which capture learners' actual interactions, can be combined with data from self-report instruments such as questionnaires or think-aloud sessions (Bannert, Reimann, & Sonnenberg, 2014; Eynon, 2013), or data from external sources like eye-tracking (Trevors, Feyzi-Behnagh, Azevedo, & Bouchet, 2016). To get a better

[*] Corresponding author. Pontificia Universidad Católica de Chile, Department of Computer Science, Avda. Vicuña Mackenna 4860, Macul, Santiago, Chile.

E-mail addresses: jjmaldonado@uc.cl (J. Maldonado-Mahauad), mar.perez@uc.cl (M. Pérez-Sanagustín), kizilcec@stanford.edu (R.F. Kizilcec), nvmorale@uc.cl (N. Morales), jmun@uc.cl (J. Munoz-Gama).

https://doi.org/10.1016/j.chb.2017.11.011
0747-5632/© 2017 Elsevier Ltd. All rights reserved.

understanding of how learners behave and learn in digital environments there is a need to explore ways to connect educational theory to data-driven methods with behavioural and self-report data (Lodge & Corrin, 2017).

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) or receive active self-regulation support from the platform (Kizilcec & Cohen, 2017). In the absence of the support and guidance that is typically available in brick-and-mortar learning environments (e.g., an instructor setting deadlines and structuring the learning process), the ability to regulate one's learning process is a critical skill to achieve personal learning objectives in a MOOC. Online learners need to determine when and how to engage with course content without any other support than the course content and structure, which can pose a challenge for many learners (Lajoie & Azevedo, 2006). Self-regulated learners are characterized 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 learning 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.

Over the last 30 years, multiple models have been developed to explain how the process of SRL develops amongst learners (Boekaerts, 1999; Borkowski, 1996; Pintrich, 2004; Winne & Hadwin, 1998; Zimmerman, 2015; Panadero, 2017). These models serve as a foundation for developing methods to study the use of SRL strategies in the learning process. They can be categorised as either component models or 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. 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). Researchers in the field of SRL have suggested that questions about measuring constructs associated with self-regulation should be seen in terms of aptitudes (for component models; Bannert et al., 2014) and events (for process models; Winne, 2010). Thus, both learner aptitudes and events contribute to a global understanding of how SRL works. On the one hand, aptitudes are essential to researching SRL since they are theoretical constructs underlying observed differences between individual learners in specific contexts such as motivational factors and epistemic beliefs (Snow, 1989). On the other hand, events are the actions that learners perform and provide touch points to map information in order to infer learners' cognitive processes (Winne, 2010).

Prior research studying SRL in MOOCs identified learner characteristics that are predictive of stronger SRL skills based on clickstream behaviour data and a survey instrument (Kizilcec, Pérez-Sanagustín, & Maldonado, 2017). This article extends these findings by leveraging process mining methods with the clickstream data collected in three MOOCs. In particular, this study focuses on the relationship between the trace data generated through the interaction of learners with the course content (video lectures and assessments) in online sessions and learners' self-reported SRL skills. Mukala, Buijs, Leemans, and Van Der Aalst (2015) found that learners interact with video-lectures, assessments and other MOOC contents week by week, identifying loopbacks, deviations and bottlenecks. The current investigation additionally incorporates data on learners' assessment submission behaviour. In this study, formal Process Mining (PM) techniques are used 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. Specifically, an analysis of learners' behaviour sequences in a MOOC from a PM perspective could enable us to understand how observed interaction sequence patterns are aligned with SRL strategies. To this end, we present the results of an exploratory sequence analysis to detect patterns in learners behaviour and combining with their SRL profile scores. The results show that learners who usually follow the sequential structure provided by the MOOC's instructional design perform more organised sessions that set them up for gaining a deeper understanding of the MOOC content (comprehensive learners), while learners that look for specific information that will help them pass the course assessments tend to be more strategic (targeting learners). Both approaches are independent of the learners' personal SRL skills and learners who take these approaches are more effective compared with sampling learners who exhibit more erratic behaviour and lower SRL skills.

In the remainder of this section, we present the theoretical background that drives the research questions of this study, the instruments employed, the PM techniques used for the analysis. The next section presents the methods and the process mining approach used to address the research questions. Section 3 reports the main results and Section 4 discusses them. Finally, Section 5 presents conclusions, limitations and implications of this work.

1.1. Self-regulated learning in online environments: interaction sequences patterns

Several studies have demonstrated a positive relationship between the use of SRL strategies in online environments and academic achievement (Broadbent & Poon, 2015; Broadbent, 2017; Richardson, Abraham, & Bond, 2012; Robbins et al., 2004; Wang, Shannon, & Ross, 2013). Most research on SRL in online environments adopts an aptitude-based approach. This research has developed various instruments to measure which SRL strategies learners use in online environments. These instruments include self-report questionnaires, think-aloud protocols (a type of interview), and learning diaries (Roth, Ogrin, & Schmitz, 2015; Wirth & Leutner, 2008). Self-report questionnaires are the most common type of assessment for SRL. They assess cognitive, metacognitive and resource management strategy use in order to identify specific learning strategies or tactics. Moreover, self-reports are feasible for large-scale assessment where observational methods are impractical (Roth et al., 2015). Some of the most established SRL questionnaires are the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich, Smith, Garcia, & McKeachie, 1991), the Academic Self-Regulated Learning Scale (ASRLS) (Magno, 2011), and the Self-Efficacy for Learning Form (SELF) (Zimmerman & Kitsantas, 2007). Subsequent studies developed novel instruments with items adapted from these established questionnaires (Roth et al., 2015). In general, these questionnaires can be used to establish aptitude-based SRL profiles for learners: for example, to distinguish between highly self-regulated and less self-regulated learners.

In recent years, there has been a boost in research to understanding SRL in online environments, in particular research that investigates SRL as a process. This is in part due to advances in digital learning environments that can record learner behaviour at a fine-grained level (e.g., information collected from a learner's interactions with the course content such as lectures or assessments). The aptitude-based approach to studying SRL has relied on questionnaires that reflect a static image of SRL. Yet SRL is a dynamic process sensitive to the specific context where learners perform a task. Thus, the process-based approach offers an opportunity to overcome some of the shortcomings of the aptitude-based approach and self-report instruments (Jovanović, Gašević, Dawson, Pardo, & Mirriahi, 2017). From this process-based perspective, SRL can be conceived as a set of events or actions that learners perform when they are studying (learning traces), rather than a description of those actions or mental states that these actions generate (Bannert et al., 2014). Recording the context of each trace is possible to obtain a representation of the performed behaviour without asking a learner about it (e.g., as with think-aloud methods) (Winne, 2013). In this sense, PM is a suitable approach for studying SRL in online environments from a process perspective. Specifically, PM facilitates the discovery of learning process models, which represent the sequence of learners' interactions with course materials (Van Der Aalst et al., 2011). It also provides robust ways of extracting, analysing and visualising learners' interaction traces (Jivet, 2016; Mukala, Buijs, & Van Der Aalst, 2015b; Romero, Cerezo, Bogarin, & Sánchez-Santillán, 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). For example, Hadwin, Nesbit, Jamieson-Noel, Code, and Winne (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, Gašević, and Hatala (2015) examined the relationship 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 70 undergraduate 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 of 53 learners, where they show that promoting social awareness strongly influenced with the micro-level processes of SRL of the learners.

This prior work demonstrates the potential of taking a PM approach to study SRL, but there are some notable limitations that need to be addressed. First, the small sample sizes and homogeneity of study participants limits the generalizability of prior findings. Second, participants were unfamiliar with the digital learning tools that were developed to assess SRL and their learning experience with these tools may not have been realistic. It is preferable to study diverse learners’ interaction traces and SRL at larger scale and in naturalistic online learning settings. Much research on SRL in online environments has been done on platforms that were either manipulated or adapted to study SRL, by adding functionalities that were associated with a self-regulated strategy (Beheshitha et al., 2015; Siadaty et al., 2016; Sonnenberg & Bannert, 2015). The use of designated learning platforms to study SRL provides greater experimental control and flexibility in measurement at the expense of external validity.

To study learning paths we consider different levels of interaction granularity by which we denote the number of events that occur over time in an interaction sequence (Bannert et al., 2014). The granularity in the interaction sequence can be studied in terms of learning trajectories that learners follow based on the content structure of a MOOC (e.g., a linear trajectory going from one week to the next). Granularity can also be studied in terms of learners’ interaction sequences with specific objects in the course, that are part of a learning activity (e.g., learning trajectories between lectures, assessments, discussion forums, etc.). Thus, the data gathered can help us gain insights into how learners engage with the course content and provide more information about tactics and strategies that might be useful when studying. Accordingly, we defined our first research question as follows:

RQ1. What are the most frequent interactions sequences of learners in MOOCs?

1.2. Self-regulated learning in MOOCs: academic performance and SRL profile

Several researchers have investigated the relationships between interaction sequences and learning outcomes using methods such as transition graphs (Hadwin et al., 2007), sequence mining to model learner behaviour (Köck & Paramythis, 2011), sequential pattern analysis (Agrawal & Srikant, 1995; Perera, Kay, Koprinska, Yacef, & Zaïane, 2009), and Markov models (Biswas, Jeong, Kinnebrew, Sulcer, & Roscoe, 2010). These methods enabled researchers to identify different learning behaviour sequences for high- and low-performing learners. Guo and Reinecke (2014) investigated the most common two-step chain interaction sequences exhibited by MOOC learners. They found that learners frequently used non-linear learning paths and performed back jumps to previous video lectures. In addition, older learners tended to plan their own learning paths, ignoring the linear course structure. Davis, Chen, Hauff, and Houben (2016b) also investigated how learners adhere to the designed paths. They used eight-step chain interaction sequences to gain insights into behavioural patterns using discrete-time Markov chains. Mukala, Buijs, and Van Der Aalst (2015a) applied PM in order to understand learning processes based on learners' interaction in a MOOC with 43,218 learners. They found that (1) successful students performed better because they followed the videos and submitted 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). Also in MOOCs, several studies have adapted and applied questionnaires to determine the level of self-regulation among MOOC learners from an aptitude perspective and their relation with academic performance (Alario-Hoyos, Estévez-Ayres, Pérez-Sanagustín, Kloos, & Fernández-Panadero, 2017; Jansen, van Leeuwen, & Janssen, 2016; Kizilcec, Pérez-Sanagustín, & Maldonado, 2016; Littlejohn, Hood, Milligan, & Mustain, 2016; Maldonado et al., 2016). These studies use a variety of instruments and methods (e.g. averaging scores for overall scales as well as a separate score, clustering to detect profiles among others) to define learners’ self-regulation profiles and how they relate with completion rates.

Although these studies report interesting results on how learners behave in a MOOC from a process perspective, they do not investigate how these learning sequences relate with SRL strategies and academic performance. The analysis of the interaction sequences patterns performed by the students can help gain insight into the types of strategies that learners use and their relative efficacy. However, as a recent study by the MOOC Research Institute points out, and to the best of our knowledge, such research is still scarce (Gašević, Kovanovic, Joksimovic, & Siemens, 2014), especially in the context of SRL and MOOCs. In order to make progress in this area, the conceptualization of 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. A recent article made a step toward this goal by analysing the relationship between self-reported SRL and actual behaviour in six MOOCs (Kizilcec et al., 2017). It 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, this prior work studied learner behaviour at the level of individual interactions (using transition probability to pass from one interaction to another) to obtain a basic process model. Yet a more fine-grained approach that considers more complex sequences is needed to understand SRL in MOOCs as a process. This brings us to our second and third research questions:

RQ2. How do the interaction sequences of learners with different academic performance differ?

RQ3. How do the interaction sequences between learners with different SRL profiles differ?

1.3. *Self-regulated learning strategies*

Self-regulated learning is a very complex process that involve both psychological and behavioural changes. Self-regulated learners are those with the ability to engage with cognitive, metacognitive, affective, and motivational processes to increase their probability of achieving their learning goals successfully (Boekaerts, 1999; Borkowski, 1996; Pintrich, 2004; Winne & Hadwin, 1998; Zimmerman, 2015). Beside these psychological processes, self-regulated learners must have the ability to initiate behavioural changes in order to take the necessary actions to achieve their learning goals and persevere until they succeed. These behavioural changes manifest as a set of actions or strategies in which learners set goals, attempt to monitor, regulate and control, guided and constrained by their goals and contextual features of the learning environment (Pintrich, 2000). Moreover, the capacity of a learner to select and adjust their learning strategy according to the requirements of the learning context is the key in order to engage in self-regulated learning (Winne, 2006). However, observing SRL strategies, even when these manifest as a set of actions and behavioural changes, entails several challenges.

The first challenge is to identify and observe behavioural changes. Even in an online environment, where learners' actions are registered, we are not capturing all the actions involved in learners' learning process. Certain strategies, such as goal setting or help seeking might be occurring beyond the learning platform. For example, we do know that MOOC learners complement their learning process with social networks (Chen, Davis, Lin, Hauff, & Houben, 2016; García-Peñalvo, Cruz-Benito, Borrás-Gené, & Blanco, 2015). However, we do know when this behaviour occurs within the learners' learning process and how this relates with SRL strategies.

The second challenge is to understand whether an observable behaviour relates to a particular SRL strategy or to more than one. For example, is possible to say that when a learner spends a study session watching video-lectures in a MOOC, it could be related to the *Study* strategy as defined by Garavalia and Gredler (2002) ("Study in a particular order"), or as *Rehearsal* as defined by Broadbent (2017) (e.g. "Learner who listens to an online lecture repeatedly"). Moreover, researchers agree that SRL is not a fixed trait, but rather a skill that can be developed through personal experiences and practice applying learning strategies (Azevedo & Cromley, 2004; Schunk, 2005; Zimmerman, 2015). This means that an observable behaviour at the beginning of the course may be related to a different strategy when it is observed at the end of the course.

To address these challenges, some researchers have made an effort to associate certain behavioural patterns with learning strategies. For example, Hadwin and Winne (2012) analysed the learning outcomes of a set of learners when applying certain strategies. They observed that individuals who apply relevant learning strategies would act more strategically and intentionally than the others, such as recalling related prior knowledge and cognitively manipulating new information to connect with their prior knowledge in order to improve retention. Jovanović et al. (2017) observed that those learners' adopting the learning strategies aligned with teachers' teaching strategy were more successful in online course.

This prior work, together with the studies identifying interaction sequences in online environments presented in Section 1.1, shed some light on how to relate observed behaviour with learning strategies. However, how MOOC learners' actions and behaviour relates with SRL strategies as defined in the theory is still unclear.

Table 1
Overview of the MOOCs in our study.

	MOOC 1	MOOC 2	MOOC 3
	($n = 497$)	($n = 2035$)	($n = 926$)
Enrolled	18653	25706	10576
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

2. Method

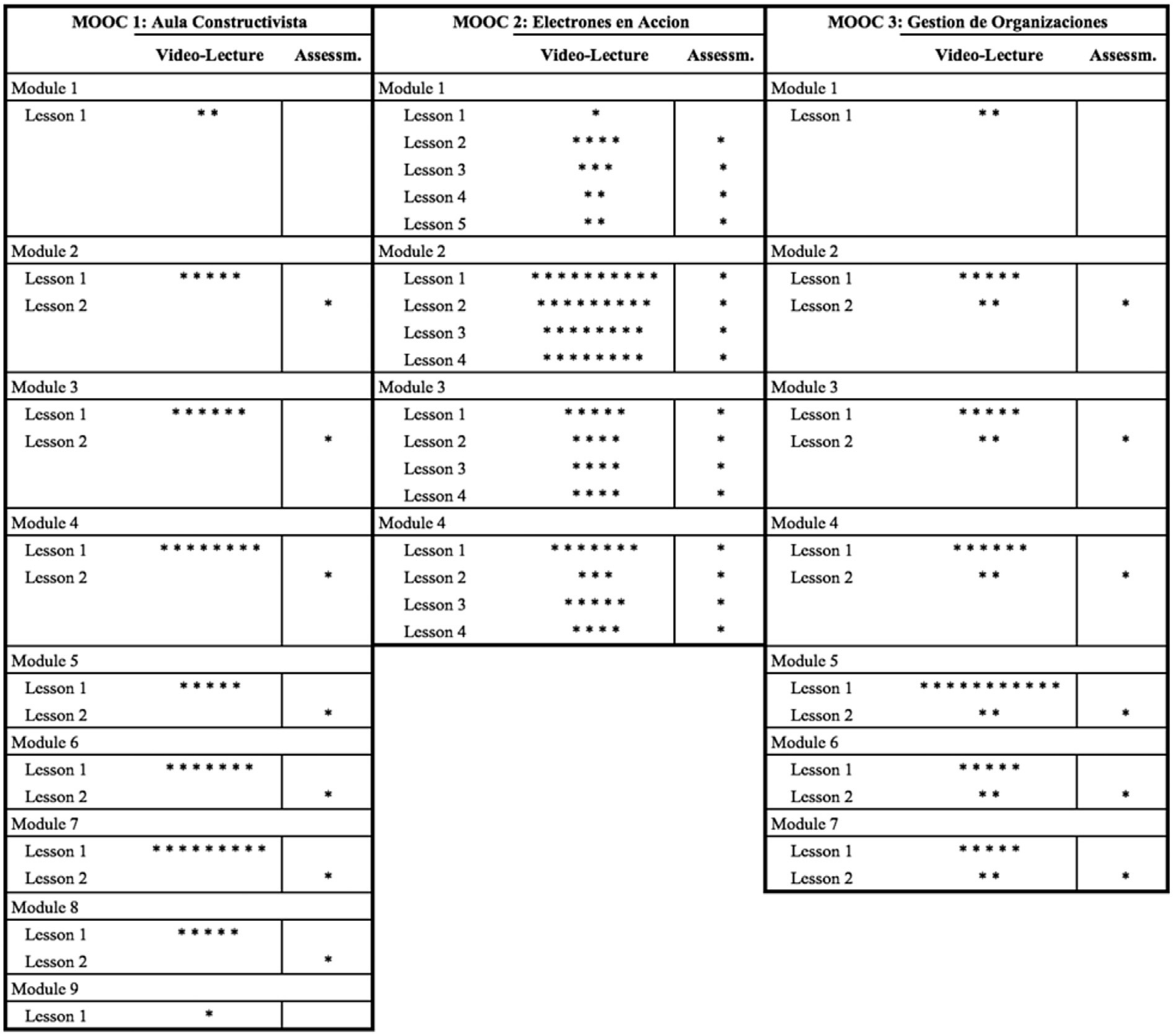
2.1. *Sample*

The final study sample comprised $N = 3458$ online learners in three different MOOCs (see 2.2. Courses). This sample is a subset of 4871 respondents who answered the initial questionnaire among the 54,935 learners who registered for the MOOCs. We excluded 1413 responses for one of the following reasons: (1) learners took the survey more than once in the same course ($n = 733$), (2) empty surveys without answers ($n = 133$), and (3) survey data could not be linked to platform data ($n = 547$). The target audiences of the three 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 April and December 2015.

2.2. *Courses*

This study encompassed three courses[1] offered by Pontificia Universidad Católica de Chile on Coursera. The courses were taught in Spanish on topics related to engineering ($n = 2035$ in final study sample), education ($n = 497$) and management ($n = 926$). The course materials were organised 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. Fig. 1 illustrates the structure of each course.

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



MOOC 1: Aula Constructivista	Video-Lecture	Assessm.
Module 1		
Lesson 1	**	
Module 2		
Lesson 1	*****	
Lesson 2		*
Module 3		
Lesson 1	******	
Lesson 2		*
Module 4		
Lesson 1	********	
Lesson 2		*
Module 5		
Lesson 1	*****	
Lesson 2		*
Module 6		
Lesson 1	*******	
Lesson 2		*
Module 7		
Lesson 1	*********	
Lesson 2		*
Module 8		
Lesson 1	*****	
Lesson 2		*
Module 9		
Lesson 1	*	

MOOC 2: Electrones en Accion	Video-Lecture	Assessm.
Module 1		
Lesson 1	*	
Lesson 2	****	*
Lesson 3	***	*
Lesson 4	**	*
Lesson 5	**	*
Module 2		
Lesson 1	**********	*
Lesson 2	*********	*
Lesson 3	********	*
Lesson 4	********	*
Module 3		
Lesson 1	*****	*
Lesson 2	****	*
Lesson 3	****	*
Lesson 4	****	*
Module 4		
Lesson 1	*******	*
Lesson 2	***	*
Lesson 3	*****	*
Lesson 4	****	*

MOOC 3: Gestion de Organizaciones	Video-Lecture	Assessm.
Module 1		
Lesson 1	**	
Module 2		
Lesson 1	*****	
Lesson 2	**	*
Module 3		
Lesson 1	*****	
Lesson 2	**	*
Module 4		
Lesson 1	******	
Lesson 2	**	*
Module 5		
Lesson 1	***********	
Lesson 2	**	*
Module 6		
Lesson 1	*****	
Lesson 2	**	*
Module 7		
Lesson 1	*****	
Lesson 2	**	*

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 that was used in prior research with MOOCs (Kizilcec et al., 2016).[3] The SRL measure consisted of 24 statements related to six SRL strategies and it was originally adapted from multiple established instruments (Barnard, Paton, & Lan, 2008; Littlejohn & Milligan, 2015; Pintrich et al., 1991; Rigotti, Schyns, & Mohr, 2008; Warr & Downing, 2000). Learners rated statements using a 5-point scale (coded from 0 to 4). The six SRL strategies that were assessed are 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 questionnaire was obtained following the same procedure as in prior work (Kizilcec et al., 2016). For each strategy, the individual score was computed by averaging ratings of corresponding statements. The SRL measure exhibited high reliability for all strategy subscales with Cronbach's alpha of at least 0.70, which is generally considered acceptable (Peterson, 1994). The SRL composite, an index of all six subscales, had very high reliability ($\alpha = 0.91$). Table 2 presents descriptive statistics for each SRL strategy and composite, also the Cronbach's α, Pearson's correlation coefficients between strategies.

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 (Fig. 2): (1) extraction - the data is extracted from the

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

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

Table 2
Descriptive statistics for each SRL strategy and composite: mean and standard deviation, Cronbach's α, Pearson's correlation coefficients between strategies, and SRL composite (averaging scores for all strategies) ($\bar{x}$). The access for the SRL questionnaire[3] is provided in the footnote.

Strategy	*M* (*SD*)	α	2.	3.	4.	5.	6.	$\bar{x}$
1. Goal Setting	3.02 (0.75)	0.86	0.70	0.46	0.57	0.46	0.29	0.78
2. Strategic Planning	3.11 (0.64)	0.73		0.60	0.65	0.58	0.31	0.84
3. Self-evaluation	3.28 (0.65)	0.79			0.62	0.60	0.24	0.73
4. Task Strategies	3.10 (0.62)	0.78				0.72	0.34	0.87
5. Elaboration	3.31 (0.63)	0.76					0.32	0.77
6. Help Seeking	2.62 (0.78)	0.75						0.58
$\bar{x}$ SRL Composite	3.06 (0.52)	0.91						

Information System data bases (Coursera in our case), (2) event log generation— the table value information is modeled in terms of event logs, defining the concepts of case (execution of a process), activities (steps of the process), and temporal order of the activities, (3) model discovery— process mining discovery algorithms are applied to the event log in order to automatically mine a process model describing the observed behaviour of the process, and (4) model analysis— the discovered process models are analysed in order to understand the observed behaviour. 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 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 organised 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 file we used in the PM algorithm. This event log is a file that stores the information on the learners' interactions within the MOOC, their SRL scores, as well as information necessary to perform the analysis such as the case id, time stamp and other resources. The first step for generating the event log 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 3.
- 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 min. This definition of session was adopted from the prior works by Kovanović et al. (2015) and Liu et al. (2015).

In addition to the interactions, the event log file included the learners' SRL scores that we obtained from the SRL self-reported questionnaire. 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 SRL scores, 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 4 shows an example of the event log generated.

2.4.3. Discovery of the model

We processed the event log with a discovery algorithm to obtain

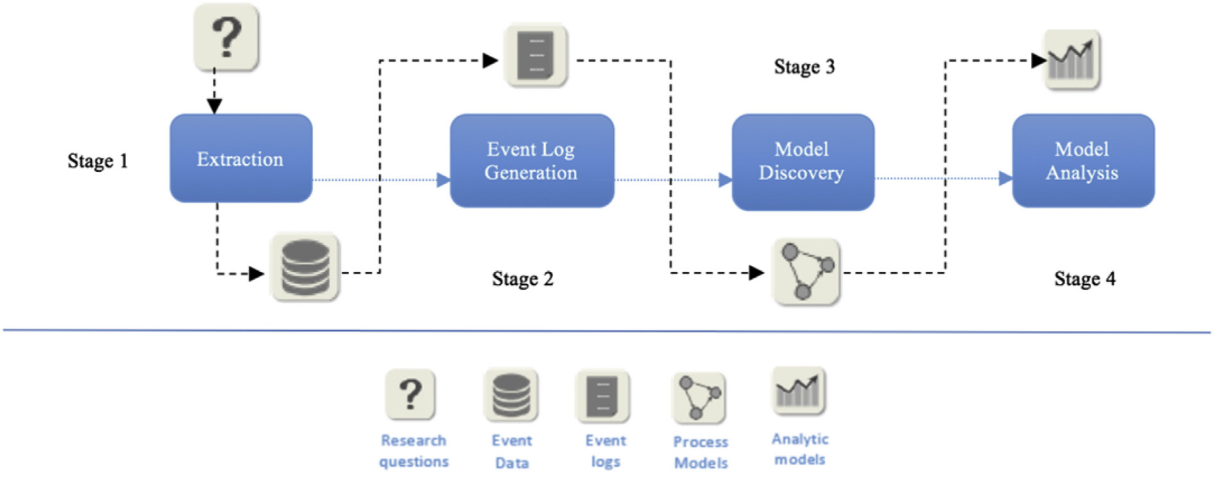


Fig. 2. Stages for the generation of the process model using PM^2 methodology. Figure adapted from Van Eck et al. (2015).

Table 3
Definitions of six interaction types with course materials to characterize consecutive learner behaviour.

Interaction	Definition
(1) Video-Lecture begin	Begin watching a video-lecture without completing it. The video-lecture was not previously completed.
(2) Video-Lecture complete	Watch a video-lecture in its entirety on the first attempt.
(3) Video-Lecture review	Go back to a video-lecture that the learner had previously watched in its entirety (not necessarily on the first attempt).
(4) Assessment try	Unsuccessful attempt to solve an assessment.
(5) Assessment pass	Successful attempt to solve an assessment for the first time.
(6) Assessment review	Go back to an assessment that was previously completed successfully (not necessarily on the first attempt).

Table 4
Example of the event log generated for the process analysis.

Case ID	Time Stamp	Interaction	SRL Scores	Course completion	Session
c7a1821f350de427f31acc92cf40b27c8a36ea9d	1451023929	Begin session	3.162	False	1
c7a1821f350de427f31acc92cf40b27c8a36ea9d	1448567431	Video-Lecture.begin	3.162	False	1
c7a1821f350de427f31acc92cf40b27c8a36ea9d	1448567737	Video-Lecture.complete	3.162	False	2
c7a1821f350de427f31acc92cf40b27c8a36ea9d	1448568139	Assessment.try	3.162	False	2
c7a1821f350de427f31acc92cf40b27c8a36ea9d	1449103918	Video-Lecture.review	3.162	False	1
011ff41dfa7cc2cf9bb89a73fd9ac1ac74eef4d3	1449104348	Assessment.pass	3.433	True	1
011ff41dfa7cc2cf9bb89a73fd9ac1ac74eef4d3	1449104694	Assessment.review	3.433	True	2
011ff41dfa7cc2cf9bb89a73fd9ac1ac74eef4d3	1449105157	End session	3.433	True	1

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 & 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 & 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. The interaction sequences were first used for an exploratory analysis and then for clustering.

As a result of applying the algorithms, we obtained a *spaghetti process* model (Fig. 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 white hexagon with a play image and a stop image inside, respectively. The interactions in Table 3 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. Fig. 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, Fig. 4 shows that the event log contains 13714 *Begin session* interactions; that is, there were 13714 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. Fig. 4 shows that the interaction sequence from *Begin-session* to *(→) Video-lecture-begin* was performed 9162 times. This means that from the 13714 sessions that were initiated, only 9162 interaction sequences were performed toward *Video-lecture-begin*.

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 firsts research questions:

RQ1. What are the most frequent interactions sequences of learners in MOOCs? 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. How do the interaction sequences of learners with different academic performance differ? 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 learners completed the course. To achieve this, we filtered the log file by completer ($n = 258$) and non-completer ($n = 3200$) status. This allowed us to observe differences between the various interaction sequence patterns. We also generated process models for completers and non-completers.

RQ3. How do the interaction sequences between learners with different SRL profiles differ? To answer this question, we use an agglomerative hierarchical clustering technique for grouping learners ($N = 3458$) based on the identified interaction sequence patterns (e.g. learning strategies). That is, we cluster learners based

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

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

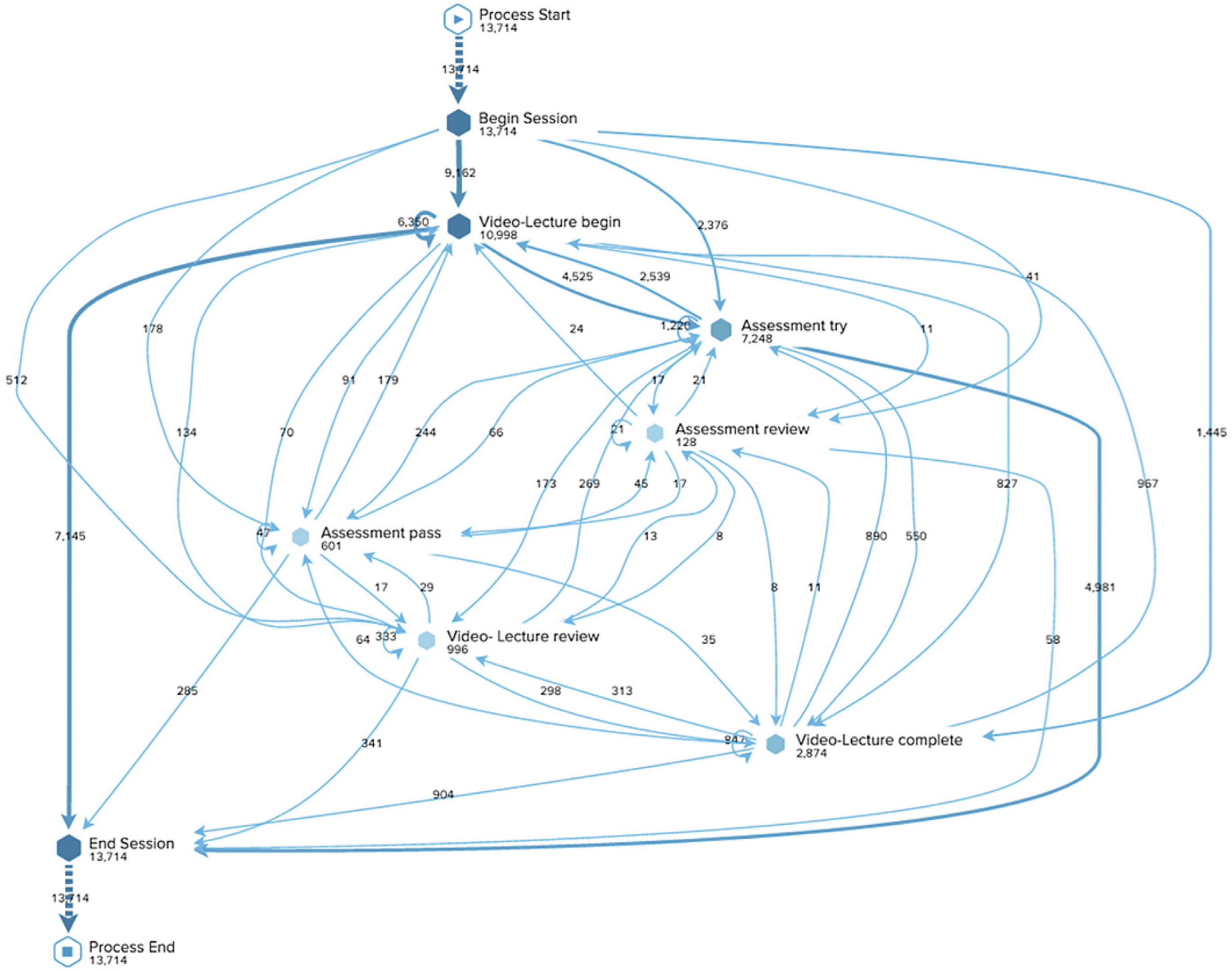


Fig. 3. *Spaghetti full process model* containing all interaction sequences of 3 MOOCs by sessions. The process model contains the six possible interactions that learners can perform with the course content like video-lecture begin, video-lecture complete, video-lecture review, assessment try, assessment pass, assessment review. Also, the process model specifies the number of sessions that start (begin session) and end (end session).

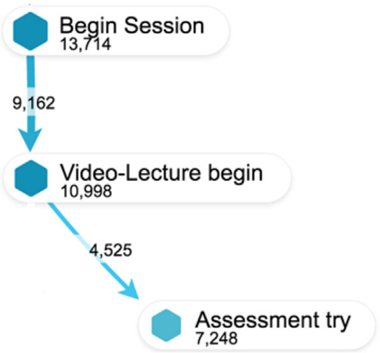


Fig. 4. Representation of interaction sequences extracted from the *spaghetti full process model*. This extract of the process model shows that the interaction sequence from Begin-session to (→) Video-lecture-begin was performed 9162 times and the interaction sequence from Video-lecture begin to (→) Assessment try was performed 4525 times. Also, the numbers under the interaction caption next to each coloured hexagon indicates the number of times the interaction caption was repeated. For this case 10998 times for Video-lecture begin interaction and 7248 times for Assessment try interaction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

on their distinct use of learning strategies. We use the scores obtained through the self-reported SRL questionnaire in order to observe how learners are distributed across the different clusters.

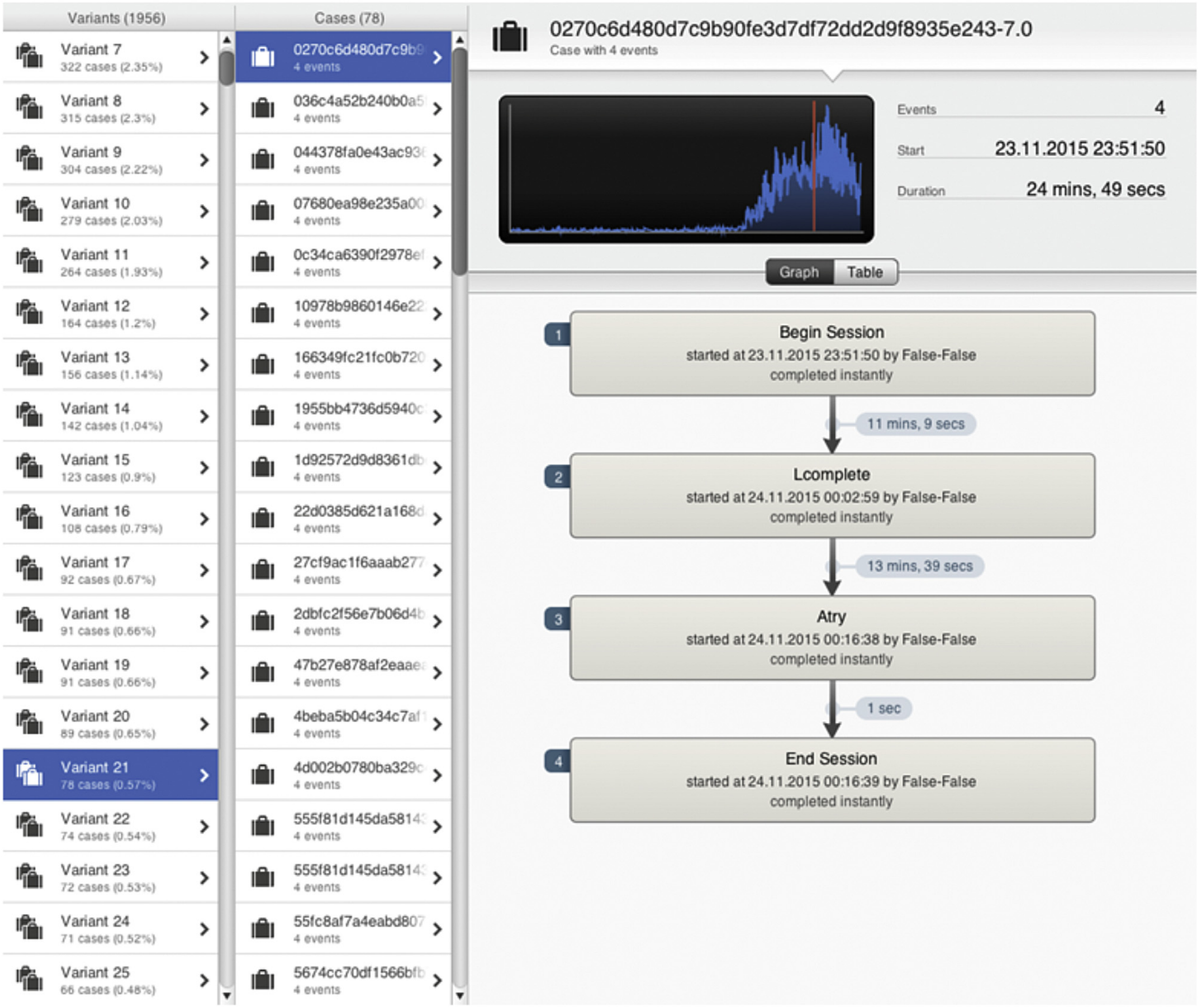
3. Results

The results section is structured around the four research questions. Additional results, tables, and supporting data are provided in the Appendix.

3.1. *What are the most frequent interactions sequences of learners in MOOCs? (RQ1)*

We generated the process model shown in Fig. 3 based on 13714 sessions. There were 1956 different types of sessions, each containing a set of interaction sequences that characterized the session. Fig. 5 shows a screenshot of the Disco software, which provides a list of the 1956 types and an overview of its related interaction sequences.

The types of sessions were ordered from the most common to the least common. The most common we assigned to a category



Variants (1956)

Variant 7
322 cases (2.35%)

Variant 8
315 cases (2.3%)

Variant 9
304 cases (2.22%)

Variant 10
279 cases (2.03%)

Variant 11
264 cases (1.93%)

Variant 12
164 cases (1.2%)

Variant 13
156 cases (1.14%)

Variant 14
142 cases (1.04%)

Variant 15
123 cases (0.9%)

Variant 16
108 cases (0.79%)

Variant 17
92 cases (0.67%)

Variant 18
91 cases (0.66%)

Variant 19
91 cases (0.66%)

Variant 20
89 cases (0.65%)

Variant 21
78 cases (0.57%)

Variant 22
74 cases (0.54%)

Variant 23
72 cases (0.53%)

Variant 24
71 cases (0.52%)

Variant 25
66 cases (0.48%)

Cases (78)

0270c6d480d7c9b9
4 events

036c4a52b240b0a5
4 events

044378fa0e43ac93
4 events

07680ea98e235a00
4 events

0c34ca6390f2978e
4 events

10978b9860146e22
4 events

166349fc21fc0b720
4 events

1955bb4736d5940c
4 events

1d92572d9d8361db
4 events

22d0385d621a168d
4 events

27cf9ac1f6aaab277
4 events

2dbfc2f56e7b06d4b
4 events

47b27e878af2eaae
4 events

4beba5b04c34c7af
4 events

4d002b0780ba329c
4 events

555f81d145da5814
4 events

555f81d145da5814
4 events

55fc8af7a4eabd807
4 events

5674cc70df1566bfb
4 events

0270c6d480d7c9b90fe3d7df72dd2d9f8935e243-7.0
Case with 4 events

Events 4
Start 23.11.2015 23:51:50
Duration 24 mins, 49 secs

1 Begin Session
started at 23.11.2015 23:51:50 by False-False
completed instantly

11 mins, 9 secs

2 Lcomplete
started at 24.11.2015 00:02:59 by False-False
completed instantly

13 mins, 39 secs

3 Atry
started at 24.11.2015 00:16:38 by False-False
completed instantly

1 sec

4 End Session
started at 24.11.2015 00:16:39 by False-False
completed instantly

Fig. 5. List of the 1956 types of the sessions obtained using Disco software. The type 21 shows 4 interactions (events) with 3 interaction sequences and the time associated with the duration of the session (variant 21).

that describes the interaction sequence pattern. For example, we analyse the first most common types of sessions and we observed that these consists in video-lecture begin interaction sequences. So, a pattern of Only video-lecture is defined. Then, we filtered the log file marking these types of sessions. After that, the procedure is repeated, identifying the rest of the sessions types that remains without mark in the log file. It was done through a python script developed. As a result, we obtained the following seven interaction sequences patterns:

1. Only Video-lecture: 2539 repetitions of the type of session
2. Only Assessment: 604 repetitions of the type of session
3. Explore: 583 repetitions of the type of session
4. Assessment try to Video-lecture: 304 repetitions of the type of session
5. Video-lecture complete to Assessment try: 78 repetitions of the type of session
6. Video-lecture to Assessment complete: 15 repetitions of the type of session
7. Others: 3 repetitions of the type of session

Those types of sessions that fit into multiple interaction sequence patterns (given that they are long and disperse) or they do not fit into any interaction sequence pattern, were classified as “Others”. The description of each interaction sequence pattern is based on whether a session only contains certain type of interaction (defined in Table 3) or whether the session contains certain type of interaction sequences between interactions that are important in the learning process (for example pass from try an assessment to a video-lecture which represents how the learner looks for missing information after not passing the assessment). Once the most common interaction patterns were extracted from the main process model (Fig. 3), we defined for each pattern a process model (Figs. 6, 8, 9, 10, 11, 12 and 13 of the Appendix), in order to observe the learner behaviour as a result of the interaction with the MOOC content in a session. We described the seven distinct interaction sequence patterns extracted by PM as follows:

(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* (Fig. 6).

(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* (Appendix - Fig. 8).

(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* (Appendix - Fig. 9).

(4) *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 (Appendix - Fig. 10).

(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* (Appendix - Fig. 11).

(6) *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* (Appendix - Fig. 12).

(7) *Others:* interaction sequence patterns that are long and disperse and they do not fit into any interaction sequence pattern mentioned before. The most common interaction sequences observed are (a) *Begin session* to various *Video-lecture-begins* then to *Assessment-try* and then *End session* (Appendix - Fig. 13).

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 disorganised 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 complete* → *Assessment-try*** (3.32%), ***Video-lecture* → *Assessment-pass*** (1.10%) and ***Others*** (2.32%) interaction sequence patterns are the least common (Appendix – Table 12 – Finding F5). These patterns help us to understand how learners behave in a session, whether they complete the course or not. In the next section, we will analyse how distinct types of learners (based on academic performance and SRL scores) perform these interaction patterns (excluding *Ohers*) that provide insights about what strategies they used throughout the course. Table 5 summarize the

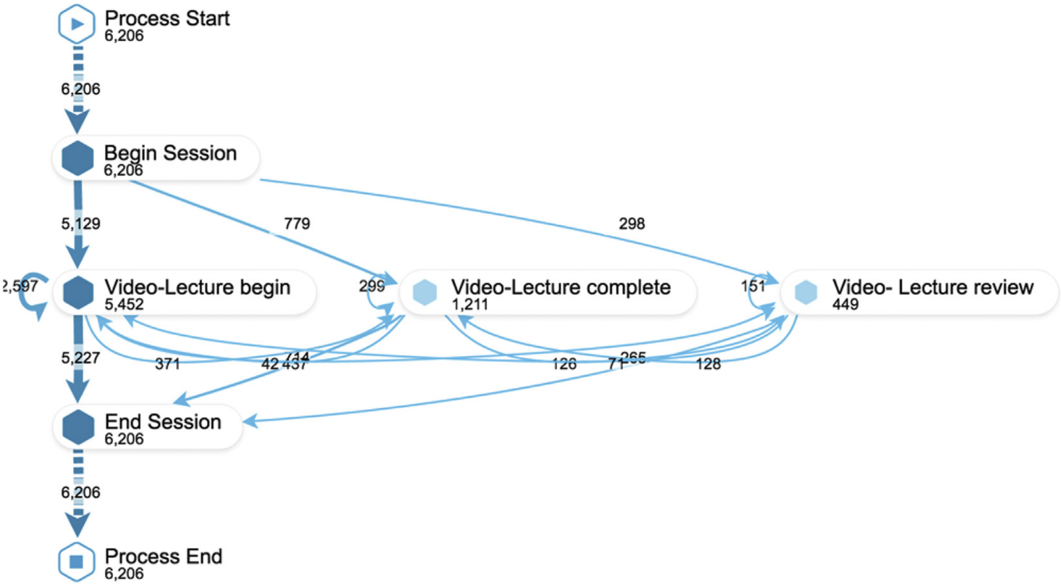


Fig. 6. Only Video-lecture interaction sequences: Process model containing interaction sequences by sessions performed with only video-lectures contents (Video-lecture begin, Video-lecture complete. Video-lecture review) being the interaction sequence Begin-session to (→) Video-lecture-begin to (→) End session the most common interaction sequence pattern. Process model generated using Celonis software.

Table 5
Proportions of the interaction sequence patterns based on the number of sessions (*N_sessions* = 13714) performed by learners in 3 MOOCs and derived from the MOOC process models.

Interaction sequence patterns	ALL 3 MOOCS		
	N_sessions	%	Learners
Only Video-lecture	6206	45.25	2495
Assessment try → Video-lecture	2960	21.58	1271
Explore	2149	15.67	1195
Only Assessment	1475	10.76	865
Video-lecture complete → Assessment try	455	3.32	358
Video-lecture → Assessment pass	151	1.10	132
Others	318	2.32	258
Total	**13714**	**100%**	-

most common patterns of interaction sequences among MOOC learners.

3.2. How do the interaction sequences of learners with different academic performance differ? (RQ2)

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. **We found that for completers were more common to perform sessions that contain more assessments than non-completers.** Completers' sessions mainly consist of: (a) taking one assessment after another (called *Only Assessment*) or (b) trying an assessment and then watching a video-lecture (called *Assessment try → Video-lecture*) or (c) watching video-lectures and trying an assessment without completing either (called *Explore*). 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 6). 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.

3.3. How do the interaction sequences between learners with different SRL profiles differ? (RQ3)

To answer this question, we started grouping learners ($N = 3458$) based on the identified interaction sequence patterns. We used agglomerative hierarchical clustering based on Ward's method. This clustering technique is advisable for detecting learner groups in online contexts (Kovanović et al., 2015). To select the optimal number of clusters we inspected the resulting dendrogram and check for different ways of cutting the tree structure, in order to obtain a minimal number of interpretable cluster explaining user behaviour (Jovanović et al., 2017). Also, we use other clustering techniques as Gaussian mixture and K-means to define the appropriate number of clusters based on the silhouette score. This lead to selecting the solution with 3 clusters as the best one (Fig. 7).

As a result, Table 7 describes the resulting clusters in terms of (1) the six identified interaction sequence patterns (we discarded “others” interaction sequence pattern as variable) used for clustering; (2) the SRL score obtained from the self-reported questionnaire; and (3) the course completion.

We have analysed similarity in the SRL profiles between each group of clusters. As a result, we did not observe statistically significant differences between Cluster 2 and 3, while we observed statistically significant differences when comparing with Cluster 1. Table 8 shows the differences between each cluster based on the SRL profile score. In the Appendix – Fig. 19 presents Box-Plot comparing SRL profiles across three clusters.

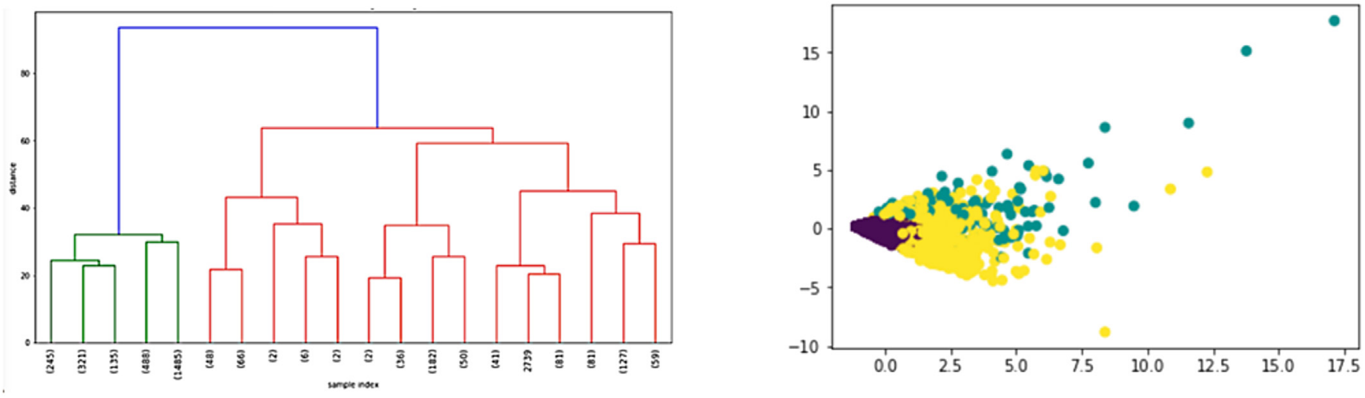
The resulting clusters indicate different kinds of learning strategies that learners have adopted while they are facing the MOOC. If we look for specifically particular differences between the different interaction sequence patterns performed by each cluster we can describe them as follows:

- Cluster 1 – Sampling learners: this cluster is composed by learners with least SRL scores compared with their counterparts. Learners in this cluster in average per session perform low number of video-lectures and in average per session perform few attempts to try to solve assessments. These learners have a low activity in the course (generally learners in this group watch just a single video-lecture or start “sample” at the beginning of the course exploring materials with the course already started).
- Cluster 2 – Comprehensive Learners: this cluster is composed by learners with a SRL scores higher than the learners in cluster 1, so they can be considered as more self-regulated (see Table 8). Learners in this cluster have developed a variety of learning strategies per session. They watched more video-lectures on average per session than learners in the other clusters. Based on the observed interaction sequences, learners in this cluster tend to follow the path that is provided by the course structure. They also invest more time watching video-lectures and therefore exhibit a higher level of engagement than learners in cluster 3. Thus, learners in cluster 2 focus on performing interaction sequence patterns in a specific order which sets them up for deeply learning the course content.

Table 6
Proportions of the interaction sequence patterns based on the number of sessions (*N_sessions* = 13714) performed in 3 MOOCs derived from the process models for Completers and Non-Completers.

Interaction sequence patterns	Completers		Non-Completers		χ2	*p*	*r*
	N_sessions	%	*N_sessions*	%			
Only Video-lecture	1253	**36.29**	4953	**48.27**	149.26	<**0.001*****	0.1043
Assessment try → Video-lecture	922	**26.70**	2038	**19.86**	71.42	<**0.001*****	0.0722
Explore	610	**17.67**	1539	**15.00**	13.94	<**0.001*****	0.0319
Only Assessment	417	**12.08**	1058	**10.31**	8.43	<**0.01*****	0.0248
Video-lecture complete → Assessment try	111	**3.21**	344	**3.35**	0.16	0.690	0.0034
Video-lecture → Assessment pass	44	**1.27**	107	**1.04**	1.26	0.262	0.0096
Others	96	**2.78**	222	**2.16**	4.34	**0.036****	0.0178
Total	**3453**	**100%**	**10261**	**100%**	-	-	

Note. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.001$.



(a) Dendrogram

(b) Scatter Plot

Fig. 7. **a)** Dendrogram obtained using agglomerative hierarchical clustering; **(b)** Scatter Plot with silhouette score = 0.5320.

Table 7
Summary statistics for the three learner clusters (sampling, comprehensive and targeting learners): median and standard deviation. For learners, completers and non-completers the number of learners and its percentage are presented under each correspondent cluster group.

	Cluster 1 – Sampling Learners	Cluster 2 – Comprehensive learners	Cluster 3 – Targeting learners
Only Video-lecture	4.67 (5.41)	22.57 (33.79)	15.72 (13.13)
Assessment try → Video-lecture	3.39 (7.09)	19.85 (18.60)	19.52 (21.42)
Explore	1.84 (3.61)	8.61 (9.40)	10.18 (11.37)
Only Assessment	0.65 (1.62)	4.18 (5.39)	4.39 (6.04)
Video-lecture complete → Assessment try	0.00 (0.00)	1.75 (3.70)	3.84 (4.95)
Video-lecture → Assessment pass	0.00 (0.00)	8.70 (6.05)	0.09 (0.80)
SRL score	3.06 (0.51)	3.12 (0.49)	3.11 (0.52)
Learners	2674 (77.32%)	124 (3.59%)	660 (19.09%)
Completers	22 (0.8%)	36 (29.03%)	200 (30.30%)
Non-Completers	2652 (99.2%)	88 (70.97%)	460 (69.70%)

Table 8
Differences between each cluster based on the SRL profile score.

Cluster #	Cluster #	t	p
2	3	0.1030	0.9179
1	2–3	−2.7333	0.0063^{***}

Note. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.001$.

- Cluster 3 – Targeting Learners: this cluster is composed of learners with similar SRL scores to those in cluster 2, which suggests that the difference in observed behaviour is not due to differences in their SRL profiles. Learners in clusters 2 and 3 also complete the course at similar rates (29% and 30% respectively). However, learners in cluster 3 watch fewer video-lectures and complete more assessments on average per session. They also tend to explore the course contents more than learners in clusters 1 and 2. These differences lead us to describe this group of learners as more strategic or goal oriented. According to Biggs (2012), strategic learners tend to focus their efforts on assessments to achieve performance-oriented objectives and exhibit less engagement overall. This interpretation is consistent with the observation that the level of engagement in cluster 3 is lower than in cluster 2.

Table 9 presents the differences found between clusters 2 and 3 in relation to interaction sequence patterns, and provide information required in order to answer the RQ3. We found statistically significant differences with significance level of 0.05 for the Only Video-lecture, statistically significant differences with significance level of 0.01 for the Video-lecture complete → Assessment try and Video-lecture → Assessment pass patterns; and statistically significant differences with significance level of 0.1 for Explore pattern, with effect sizes (r) ranging from small (Only Video-lecture, Explore); medium (Video-lecture complete → Assessment try) and big (Video-lecture → Assessment pass).

4. Discussion

4.1. RQ1. Identifying the most frequent interactions sequences of learners in MOOCs

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) trying 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 trying an assessment; and (6) watching video-lectures and trying an assessment without completing either. The extracted patterns can be interpreted as manifestations of specific learning strategies (Winne, 2013) and thus it is possible to link behavioural patterns to learning strategies. However, these patterns are only a first step towards understanding how learners self-regulate in a MOOC. More research is needed to refine and extend the identified patterns, for instance by incorporating more information such as the amount of time spent in each interaction sequence. This type of information would shed more light on how much effort learners invest in applying a particular strategy. Moreover, the current

Table 9
Comparisons respect interaction sequence patterns performed between Comprehensive and Targeting learners.

	Comprehensive learners	Targeting learners	t	p	r
Only Video-lecture	22.57	15.72	2.2276	0.0276**	0.1917
Assessment try → Video-lecture	19.85	19.52	0.1788	0.8583	0.0129
Explore	8.61	10.18	1.6393	0.100*	0.1159
Only Assessment	4.18	4.39	0.3880	0.6984	0.0284
Video-lecture complete → Assessment try	1.75	3.84	5.4396	<**0.001***	0.3476
Video-lecture → Assessment pass	8.70	0.09	15.8244	<**0.001***	0.6859

Note. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.001$.

findings could be complemented with a qualitative study focusing on why and how learners choose to use specific learning strategy.

4.2. RQ2-RQ3: differences in the identified interaction sequences between learners with different academic performance and different SRL profile in MOOCs

We found that learners who completed the course exhibited different interaction patterns than those who did not complete it. Unsurprisingly, completers were more engaged with assessments than non-completers. Going deeper, we were able to identify three types of learners in terms of their behavioural and SRL characteristics: (1) Comprehensive Learners, who have a high SRL profile, tend to follow the sequential structure of the course materials in the MOOC (i.e., guided by instructional design), and engage in more organised sessions that allow them to gain a deeper understanding of the content; (2) Targeting Learners, who also have a high SRL profile but who strategically seek out specific information to pass the course assessments; and (3) Sampling Learners, who have a low SRL profile, tend to behave in irregular ways, and are the least likely to complete the course. This clustering is consistent with findings in prior research. Kizilcec, Piech, and Schneider (2013) originally identified four clusters of prototypical MOOC learners: Completing, Disengaging, Auditing, and Sampling Learners. In comparison, Sampling Learners explore parts of the course, while Comprehensive and Targeting Learners appear to be two types of Completing Learners who may pursue different goals: deep learning and certification, respectively. Beheshitha et al. (2015) examined learners' cognitive SRL strategies while using the nStudy tool and found differences between Deep and Surface Learners that partly map onto the current distinction between Comprehensive and Targeting Learnings. Relatedly, Kovanović et al. (2015) identified three profiles and interpreted them in terms of deep versus surface approaches to learning and performance versus mastery achievement goal orientations.

4.3. Relating identified interaction sequences to SRL strategies described in the literature

We have identified six interaction patterns based on the most frequent interaction sequences observed from the trace data. Two of these interaction patterns (*Only Video-lecture* and *Only Assessment*) are composed of either interaction with video-lectures or with assessments. The other four patterns (*Assessment try→Video-lecture, Video-lecture→Assessment pass, Video-lecture-complete→Assessment try, Video-lecture-complete→Assessment try and Explore*) consist of combinations of interactions including video-lectures and assessments. We attempt to reconcile the identified behavioural patterns with SRL strategies that are established in the literature. Table 10 summarizes the relationship between these observed patterns and SRL theory.

We were able to associate each interaction sequence pattern to one or more theory-based SRL strategies. First, the *Only Video-lecture* interaction pattern was associated with three SRL strategies in the literature: **studying** (Garavalia & Gredler, 2002), **rehearsing** (Broadbent, 2017), and **repeating** (Sonnenberg & Bannert, 2015). All three are cognitive SRL strategies in which learners invest time to better understand a particular idea or knowledge component in the course. Interpretation of this interaction pattern could be enriched with additional information from external resources (e.g., capturing trace data outside the platform). This would provide more insight into whether learners use organizational SRL strategies, such as note taking, creating concept maps, or using other means to make sense of the content. As Veletsianos, Reich, and Pasquini (2016) state, "*automatically collected data by learning platforms does not necessarily offer a comprehensive and complete representation of learners' behaviour.*" Second, the *Only Assessment* interaction pattern was associated with two cognitive SRL strategies: **elaboration** (Weinstein, Acee, & Jung, 2011) and **evaluation** (Sonnenberg & Bannert, 2015). This interaction pattern was most frequently observed among the strategic Targeting Learners who are likely to complete the course (cf. Tables 6 and 9). Information about this interaction pattern could be complemented with additional information about the actions learners perform to connect the new information to their prior knowledge, and to gain more insight into whether they process information in a deep or superficial way. Third, the *Assessment try→Video-lecture* interaction pattern, which was most common among completers (cf. Table 6), was associated with **help-seeking** (Corrin, de Barba, & Bakharia, 2017; Karabenick & Dembo, 2011). Help seeking in online environments can mean that a learner looks for human help through forums, chats, or other online communication mechanisms (Broadbent & Poon, 2015). However, help can also be sought from course-internal resources (e.g. video-lectures, forums, assessments) or external resources (digital or physical material outside the platform). Thus, to better understand applications of this strategy, there is a need to collect qualitative data from interviews or focus groups asking learners about their help-seeking behaviour in MOOCs. Fourth, the *Video-lecture→Assessment pass* interaction pattern, which was most common among Comprehensive Learners (cf. Table 9), was associated with the **reviewing record strategy** (Zimmerman & Pons, 1986). This interaction pattern may reflect MOOC teachers' and instructional designers' intentions for how learners should proceed in the course: first watch a video-lecture and then pass an assessment. Fifth, the *Video-lecture-complete→Assessment* interaction pattern, which was most common among Targeting Learners (cf. Table 9), was associated with **self-evaluation** (Zimmerman & Pons, 1986). This is a metacognitive SRL strategy that has learners tracking themselves and checking their progress in the course. With the appropriate feedback, it would be possible to develop a mechanism of self-monitoring that could help learners regulate how they approach the learning process. Finally, the *Explore* interaction pattern was associated with **task exploration** (Van Der Linden, Sonnentag, Fresen, & Van Dyck, 2010). This pattern was mainly performed by Targeting Learners (cf. Table 9) and it appeared to be a strategic behaviour, for instance, switching

Table 10
Connecting Theory-based SRL strategies to patterns from observed learning behaviour.

Interaction Pattern	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 and 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 and 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 and 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 (Corrin et al., 2017; Karabenick & Dembo, 2011). 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→Assessment pass	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→Assessment pass*). Interaction sequences (a), (b), (c) and (d) could be associated with the ***Reviewing record*** SRL strategy described by Zimmerman and 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 attempts to 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 and 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, Frese, and Van Dyck (2010) (e.g. "The task exploration strategies performed in order to obtain more information and plan for learning a new computer program").

between video-lectures and assessments without completing them to investigate how the topics and the materials are organised.

Based on this preliminary pattern-strategy mapping, we found that Comprehensive Learners tended to use rehearsal, repeating, studying, reviewing record, and self-evaluation SRL strategies. Moreover, these learners tended to go back and forth over the course content to review video-lectures before and after completing an assessment, a behaviour that could be a form of *cognitive retrieval practice* (Davis, Chen, Van Der Zee, Hauff, & Houben, 2016a; Johnson & Mayer, 2009; Roediger & Butler, 2011). Conversely, Targeting Learners tended to use evaluation, elaboration, and task-exploration SRL strategies. These learners acted strategically, since they sought out specific information that would help them pass course assessments. Both Comprehensive and Targeting Learners tended to use a form of help-seeking SRL strategy.

5. Conclusions, limitations and implications

5.1. Conclusions

This article presents an empirical study on how to relate observed behaviour in a digital learning environment to established theoretical accounts of relevant learning processes. The study combines an aptitude-based approach with a process-based approach to investigate SRL strategies in MOOCs by relying on both a self-report instrument and process mining of behavioural learner data. There are four primary contributions of this research to advance the science and practice of learning:

1. Identification of the six most frequent interaction sequence patterns that learners exhibit in a MOOC;
2. Differentiation of interaction sequence patterns between learners with different course performance: completing learners interacted more frequently with assessments than those who do not complete;
3. Identification of three learner profiles based on their observed interaction sequence patters and informed by prior research on clustering learner behaviour: Comprehensive Learners who follow the "expected" sequential structure of the MOOC; Targeting Learners who seek out information required to pass assessments, and Sampling Learners who behave in irregular and unstructured ways; and
4. Association of observed interaction sequence patterns with SRL strategies established in SRL theory.

5.2. Limitations

The findings of this study are subject to some limitations posed by the nature of the data and methodological choices. First, we conducted an observational field study with automatically recorded behavioural records and data collected from an optional survey. The observations thus occurred in an actual learning environment, which is a relatively uncontrolled research setting. Prior work on SRL and learning processes that was conducted in online environments utilized research platforms developed or adapted to support SRL, for instance by adding functionalities directly associated with a self-regulation strategy (Beheshitha et al., 2015; Sonnenberg & Bannert, 2015). Aside from previous studies conducted with courses in traditional higher education settings (Jovanović et al., 2017; Lust, Elen, & Clarebout, 2013), this is to our best knowledge the first process-based study of SRL in MOOCs. Field studies in MOOCs typically yield higher external validity for lower control over the research process. For example, the optional nature of the self-report SRL instrument can raise concerns about self-selection bias, because the survey was used as a basis for including learners in the final study sample. This implies that participants in the study tended to be more motivated than the average learner enrolled in the courses.

Second, we made a number of methodological choices in this study that may have influenced the results. For example, we computed the session time based on an inactivity threshold of 45 min and we only studied learners' interactions with two learning resources in the course (video-lectures and assessments), excluding interactions on the discussion forums (this decision was made because hardly any forum interactions occurred). We highlight three methodological choices in the analysis that may have influenced our findings. First, like in any data mining or machine learning context one cannot assume to have seen all possibilities in the 'training material' (Van Der Aalst, 2016). Processes typically allow for an exponential or even infinite number of different patterns. It is therefore unrealistic to assume that every possibility is represented in the dataset. Instead, the data is considered a sample of learners' potential and observable behaviour (Bose, Mans, & Van Der Aalst, 2013). We worked solely with data from Coursera in this study. In a future project, we plan to perform the same analysis on other platforms to understand the extent to which the present findings are contextually bounded to the affordances of the learning environment. Recent evidence suggests the importance of contextual factors on learner behaviour, but it has not been analysed on a process level to date (Conole, 2015). Second, complex multidimensional and multi-granular data needs to be 'flattened' in order to be represented by simple process models (Van Der Aalst, 2016). We attempted to retain a fine level of granularity in the behavioural models, but other levels of granularity are also possible. Finally, process analysis is, by definition, restricted by the expressive power of the process modeling language (Van Der Aalst, 2011). If the modeling language cannot represent something, then it cannot be observed, resulting in representational bias. The simple process maps used to illustrate the interaction patters in this study were closely aligned with the analysis of SRL, but alternative process modeling notations with more complex patterns could also be possible. However, the discovery of more complex patterns poses additional challenges. Overall, we used transparent definitions of events and described our methodology in detail to provide the necessary accuracy to make this research reproducible. Our hope is that this article can serve as a reference point for other researchers who would like to analyse their courses using a PM approach combined with self-report data to advance our scientific understanding of how individuals learn in MOOCs.

We argue that despite these limitations the article advances our understanding of SRL in online learning. The findings complement results in prior work that was conducted in more controlled learning environments by contributing an account that focuses on MOOCs, which provide an open learning environment with a highly diverse learner population. SRL is critical for academic success in MOOCs and other settings with low levels of external guidance. Most MOOCs do not provide additional support to learners beyond the course content. This can make it difficult for learners to follow the course and achieve their learning objectives compared with a traditional classroom environment, where a teacher can support struggling learners. Online learners need to determine when and how to engage with course content without much external guidance.

5.3. Theoretical, practical and methodological implications

This exploratory study offers theoretical, practical and methodological implications and contributes to the academic discourse on how to study SRL (Azevedo, 2015; Gašević, Jovanović, Pardo, & Dawson, 2017; Winne & Jemison-Noel, 2003).

Theoretical implications. The diversity in theoretical accounts of SRL that have developed over the last 30 years has created some level of confusion with regards to SRL terminology and definitions in the literature (Winters, Greene, & Costich, 2008; Panadero, 2017). When researchers select a theoretical SRL model such as the socio-cognitive model by Zimmerman (1998), the SRL framework by Pintrich (2004), or the information processing model by Winne and Hadwin (1998), they choose a specific research path that shapes which kind of data they gather from inside or outside of the learning platform and how they analyse it (Veletsianos et al., 2016). Thus, one of the challenges of working with a non-unified SRL model is that there are multiple interpretations and analyses relevant to SRL strategies or phases of the SRL processes. In this article, we propose that certain assumptions about SRL strategies arising from the current SRL models can be studied using PM by extracting macro-level patterns from actual behavioural data. The study of theoretical assumptions at the micro-level would require other analyses and other information that SRL models are not capturing at the moment (Bannert et al., 2014). For example, the granularity in the interaction sequences can be studied in terms of learning trajectories that learners follow based on how MOOCs structure their contents (e.g. linear trajectory week by week; cf. Davis et al., 2016b). It also can be studied in terms of learners' observed interaction sequences with learning activities in the course (e.g., learning trajectories between video-lectures, assessments, forums, etc.). From this perspective, our theoretical contribution in this paper goes in the effort to relate observable processes to theory-based SRL strategies.

Practical implications. The findings of this research can also inform the design of learning environments in several ways. New insight into the behavioural signatures of SRL strategies from this research can support the implementation of accurate SRL detection systems in learning environments. These systems could help learners monitor their own use of SRL strategies in the learning process and even support social learning by comparing how their strategy use differs from that of other learners (see e.g., Davis et al., 2017). The current findings inform the development of SRL interventions for promote student learning, such as prompts based on the identified learning strategies. Moreover, the process-based findings on effective SRL strategies can inform the engineering of adaptive SRL support systems for online courses. This type of system would guide learners based on their goals and prior behaviour to take specific regulatory actions to foster motivation and robust learning outcomes. Finally, it may be possible by leveraging the detected interaction patterns to identify sections of the course that pose a self-regulatory challenge and induce high cognitive load. A process-based analysis could help identify these places in the course and minimize their negative impact by modifying the instructional design.

Methodological implications. This study demonstrates how data obtained from parsing and process mining trace data can effectively complement data obtained from self-report measures. This mixed-methods approach enables researchers to check if what learners have self-reported is consistent with their actual course behaviour. The actual MOOC platforms provide trace data as a result of the learner interaction with the course content. These platforms register a large quantity of trace data, which we filtered and processed to define events that are relevant to this study's research questions. The raw trace data is not easy to interpret and significant effort is required to parse it (i.e., extraction and cleaning methods are required before one can use the trace data). MOOC platforms could provide enriched semantic data that allow researchers to extract more interpretable information about the types of interactions that learners perform on the MOOC platform. The selected level of granularity is an important factor in the analysis of SRL from trace data. 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 its relevant phases. Different levels of granularity provide different information about SRL processes, if SRL strategies can be observed directly on the platform. In addition, common units of measurement could help researchers compare the SRL processes between MOOC platforms (e.g., defining a session as a period of time in which a learner goes through a self-regulation process, or defining the time frame as a week or the entire course duration). Moreover, common log files would make comparisons easier for researcher, and facilitate use and re-use of data across research projects to improve the learning experiences.

The vast pool of methods to choose from for the study of educational practices can present a challenge. Baker and Inventado (2014) provide a constructive synthesis of methods (e.g. classification, regression, sequence pattern mining, clustering, etc.) and emerging methodological trends in the Educational Data Mining (EDM) and Learning Analytics (LA) communities. These communities share a common interest in data-intensive approaches to education research, but according to Baker and Inventado there is an important difference: "*While LA has a relatively focus on human interpretation of data and visualization, EDM has a relatively greater focus on automated methods*". The LA and EDM communities differ primarily in their focus, research questions and the eventual use of models (Siemens & d Baker, 2012) rather than in their methodologies. In this article, in order to take advantage of big educational data, we examine how different methodological approaches from both communities can help to answer different research questions. We propose that certain assumptions about SRL strategies arising from current SRL models can be studied using PM techniques, which could extend the framework created by Baker and Inventado (2014). PM aims to quantify latent processes by extracting macro-level patterns from longitudinal log data and interpreting them as series of activities an individual engages in. It is a viable approach for process model discovery (i.e. models that accurately represent behaviour) as well as conformance checking and enhancement (i.e. finding deviations between observed and modeled behaviour, and improving the model with more data on observed behaviour) (Bogarín, Cerezo, & Romero, 2017).

Insights from process mining could be further enriched with eye-tracking data to better understand learners' cognitive learning processes, or with data on what learners do on their computer outside of the course platform. The latter could be achieved with a learning plug-in that extends the data collection to include actions that learners perform in other browser tabs, such as searching for new materials that complement the course contents, or exploring other websites with relevant information, or even engaging with irrelevant information. These additional sources of data can be correlated with learning outcomes and be used to quantify effective applications of SRL strategies. Harnessing learners' detailed behavioural records can provide an objective longitudinal account of learning and enable real-time support and feedback in ways that questionnaire data never could. This can accelerate efforts to build tools that promote SRL in MOOCs. It is this intersection of diverse data sources and experimentation that warrants much future research. *"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."* (Kizilcec & Brooks, 2017). In conclusion, although MOOCs are content-oriented settings, it would be beneficial to additionally consider them from a process-oriented perspective to be able to adapt them to support learners' SRL needs.

Compliance with ethical standards

Informed consent was obtained for use survey-instrument with human subjects.

Conflicts of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was funded by the CONICYT/DOCTORADO NACIONAL 2016 [grant No.21160081] and supported by FONDECYT (Chile) projects initiation [grant No.11150231], [grant No.11170092], the MOOC-Maker Project [grant No.561533-EPP-1-2015-1-ES-EPPKA2-CBHE-JP], the LALA Project [grant No.586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP], Ph.D. Student Fellowships University of Cuenca – Ecuador [grant No. 007-2015], and the Celonis Academic Alliance and Fluxicon Academic Initiative.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.chb.2017.11.011.

References

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the eleventh international conference on data engineering* (pp. 3–14).

Alario-Hoyos, C., Estévez-Ayres, I., Pérez-Sanagustín, M., Kloos, C. D., & Fernández-Panadero, C. (2017). Understanding learners' motivation and learning strategies in MOOCs. *The International Review of Research in Open and Distributed Learning, 18*(3).

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. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist, 50*(1), 84–94.

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology, 96*(3), 523.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61–75). New York: Springer.

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).

Biggs, J. (2012). What the student does: Teaching for enhanced learning. *Higher Education Research and Development, 31*(1), 39–55. https://doi.org/10.1080/07294360.2012. 642839.

Biswas, G., Jeong, H., Kinnebrew, J. S., Sulcer, B., & Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning, 5*(02), 123–152.

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.

Bogarín, A., Cerezo, R., & Romero, C. (2017). *A survey on educational process mining.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.

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). In *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* (Vol. 8, pp. 13–25).

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.

Chen, G., Davis, D., Lin, J., Hauff, C., & Houben, G.-J. (2016). Beyond the MOOC platform: Gaining insights aboutLearners from the social web. *Proceedings of the 8th ACM Conference on Web Science*. https://doi.org/10.1145/2908131.2908145. WebSci'16, 15e24.

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.

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).

Daradoumis, T., Bassi, R., Xhafa, F., & Caballe, S. (2013). A review on massive e-learning (MOOC) design, delivery and assessment. In *2013 eighth international conference on P2P, parallel, grid, cloud and internet computing* (pp. 208–213). https://doi.org/10.1109/3PGCIC.2013.37.

Davis, D., Chen, G., Hauff, C., & Houben, G. J. (2016b). Gauging MOOC learners' adherence to the designed learning path. In *EDM* (pp. 54–61).

Davis, D., Chen, G., Van Der Zee, T., Hauff, C., & Houben, G. J. (2016a). 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).

Davis, D., Jivet, I., Kizilcec, R. F., Chen, G., Hauff, C., & Houben, G. J. (2017). Follow the successful crowd: Raising MOOC completion rates through social comparison at scale. In *LAK* (pp. 454–463).

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? Learning,. *Media and Technology, 38*(3), 237–240. https://doi.org/10.1080/17439884.2013.771783.

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.

García-Peñalvo, F. J., Cruz-Benito, J., Borrás-Gené, O., & Blanco, Á. F. (2015). Evolution of the conversation and knowledge acquisition in social networks related to a MOOC course. In *International conference on learning and collaboration technologies* (pp. 470–481).

Gašević, D., Jovanović, J., Pardo, A., & Dawson, S. (2017). Detecting learning strategies with analytics: Links with self-reported measures and academic performance. *Journal of Learning Analytics, 4*(1).

Gašević, 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.

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).

Guo, P. J., & Reinecke, K. (2014). Demographic differences in how students navigate through MOOCs. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 21–30).

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.

Hadwin, A. F., & Winne, P. H. (2012). Promoting learning skills in undergraduate students. *Enhancing the Quality of Learning: Dispositions, Instruction, and Mental Structures*, 201–227.

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 OfComputing in Higher Education*, 1–22. https://doi.org/10.1007/s12528-016-9125-x. Springer.

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, 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.

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

Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education, 33*, 74–85.

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

Kizilcec, R. F., & Brooks, C. (2017). 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., & Cohen, G. L. (2017). Eight-minute self-regulation intervention raises educational attainment at scale in individualist but not collectivist cultures. *Proceedings of the National Academy of Sciences*, 201611898.

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2016). Recommending self-regulated learning strategies does not improve performance in a MOOC. In *Proceedings of the third ACM conference on Learning@Scale*.

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Computers & Education, 14*, 18–33.

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

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.

Köck, M., & Paramythis, A. (2011). Activity sequence modelling and dynamic clustering for personalized e-learning. *User Modeling and User-adapted Interaction, 21*(1), 51–97.

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).

Lajoie, S. P., & Azevedo, R. (2006). Teaching and learning in technology-rich environments. *Handbook of Educational Psychology, 2*, 803–821.

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)*.

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., Brown, Hartnett, Brown, M., Hartnett, M., et al. (2012). In *Future challenges, sustainable futures. Proceedings ascilite Wellington 2012*.

Lust, G., Elen, J., & Clarebout, G. (2013). Students' tool-use within a web enhanced course: Explanatory mechanisms of students' tool-use pattern. *Computers in Human Behavior, 29*(5), 2013–2021. https://doi.org/10.1016/j.chb.2013.03.014.

Magno, C. (2011). Validating the academic self-regulated learning scale with the motivated stategies for learning questionnaire (MSLQ) and learning and study strategies inventory (LASSI). *The International Journal of Educational and Psychological Assessment, 7*(2), 56–73. https://doi.org/10.1037/t09161-000.

Maldonado, J. J., Palta, R., Vázquez, J., Bermeo, J. L., Pérez-Sanagustín, M., & Munoz-Gama, J. (2016). Exploring differences in how learners navigate in MOOCs based on self-regulated learning and learning styles: A process mining approach. In *Computing conference (CLEI), 2016 XLII Latin American* (pp. 1–12).

Mukala, P., Buijs, J. C., Leemans, M., & Van Der Aalst, W. M. (2015). Learning analytics on Coursera event data: A process mining approach. In *SIMPDA* (pp. 18–32).

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.

Panadero, E. (2017). A review of self-regulated Learning: Six models and four directions for research. *Frontiers in Psychology, 8*.

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering, 21*(6), 759–772.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research, 21*(2), 381–391.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation*. San Diego: Academic Press, 451e502. https://doi.org/10.1016/B978-012109890-2/50043-3.

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

Pintrich, P. R., Smith, D., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the motivated strategies for learning questionnaire (technical report 91-B-004)*. The Regents of the University of Michigan.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353.

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.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*(2), 261–288. https://doi.org/10.1037/0033-2909.130.2.261.

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, 61*, 224–236.

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.

Schunk, D. H. (2005). Self-regulated learning: The educational legacy of Paul R. Pintrich. *Educational Psychologist, 40*(2), 85–94.

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.

Siemens, G., & d Baker, R. S. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252–254).

Snow, R. E. (1989). *Aptitude-treatment interaction as a framework for research on individual differences in learning*.

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.

Trevors, G., Feyzi-Behnagh, R., Azevedo, R., & Bouchet, F. (2016). Self-regulated learning processes vary as a function of epistemic beliefs and contexts: Mixed method evidence from eye tracking and concurrent and retrospective reports. *Learning and Instruction, 42*, 31–46.

Van Der Aalst, W. (2011). *Process mining: Discovery, conformance and enhancement of business processes*. Springer Science & Business Media.

Van Der Aalst, W. (2016). *Process mining: Discovery, conformance and enhancement of business processes (second edi.). Book*. Springer Science & Business Media.

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). PMˆ 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.

Wang, C.-H., Shannon, D. M., & Ross, M. E. (2013). Students' characteristics, self-regulated learning, technology self-efficacy, and course outcomes in online learning. *Distance Education, 34*(3), 302–323.

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. (2006). How software technologies can improve research on learning and bolster school reform. *Educational Psychologist, 41*(1), 5–17.

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist, 45*(4), 267–276.

Winne, P. H. (2013). Learning strategies, study skills, and self-regulated learning in post- secondary education. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory and research* (pp. 377–403). Netherlands: Springer. https://doi.org/10.1007/978-94-007-5836-0_8.

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. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Mahwah, NJ, U.S: Lawrence Erlbaum Associates Publishers.

Winne, P. H., & Jamieson-Noel, D. L. (2003). Self-regulating studying by objectives for learning: Students' reports compared to a model. *Contemporary Educational Psychology, 28*, 259–276. https://doi.org/10.1016/S0361-476X(02)00041-3.

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. (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. (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.

Zimmerman, B., & Kitsantas, A. (2007). Reliability and validity of self-efficacy for learning form (SELF) scores of college students. *Zeitschrift Für Psychologie, 215*(3), 157–163.

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.