Adaptation of a Process Mining Methodology to Analyse Learning Strategies in a Synchronous Massive Open Online Course

Jorge Maldonado-Mahauad[1](✉), Carlos Alario-Hoyos[2], Carlos Delgado Kloos[2], and Mar Perez-Sanagustin[3]

[1] Department of Computer Science, Universidad de Cuenca, Cuenca, Ecuador
jorge.maldonado@ucuenca.edu.ec

[2] Department of Telematics Engineering, Universidad Carlos III de Madrid, Getafe, Spain
{calario,cdk}@it.uc3m.es

[3] Institute de Recherche Informatique de Toulouse, Université de Toulouse, Toulouse, France
mar.perez-sanagustin@irit.fr

Abstract. The study of learners' behaviour in Massive Open Online Courses (MOOCs) is a topic of great interest for the Learning Analytics (LA) research community. In the past years, there has been a special focus on the analysis of students' learning strategies, as these have been associated with successful academic achievement. Different methods and techniques, such as temporal analysis and process mining (PM), have been applied for analysing learners' trace data and categorising them according to their actual behaviour in a particular learning context. However, prior research in Learning Sciences and Psychology has observed that results from studies conducted in one context do not necessarily transfer or generalise to others. In this sense, there is an increasing interest in the LA community in replicating and adapting studies across contexts. This paper serves to continue this trend of reproducibility and builds upon a previous study which proposed and evaluated a PM methodology for classifying learners according to seven different behavioural patterns in three asynchronous MOOCs of Coursera. In the present study, the same methodology was applied to a synchronous MOOC on edX with $N = 50,776$ learners. As a result, twelve different behavioural patterns were detected. Then, we discuss what decision other researchers should made to adapt this methodology and how these decisions can have an effect on the analysis of trace data. Finally, the results obtained from applying the methodology contribute to gain insights on the study of learning strategies, providing evidence about the importance of the learning context in MOOCs.

Keywords: Learning analytics · Learning behaviour · Learning strategies · Process mining · Massive open online courses

1 Introduction

One of the greatest challenges of Massive Open Online Courses (MOOC) learners is to be able to self-direct and self-regulate their learning process and adjust their strategies according to the particular context in order to achieve their learning objectives

J. Herrera-Tapia et al. (Eds.): TICEC 2022, CCIS 1648, pp. 117–136, 2022.
https://doi.org/10.1007/978-3-031-18272-3_9

[29]. In the past years, and due to the massive amount of data collected from MOOC platforms, several researchers in the Learning Analytics (LA) community have focused on the analysis of learners' trace data to unveil their learning strategies and propose new classifications accordingly [19, 22]. Several methods and techniques have been applied to analyse these trace data, such as unsupervised machine learning techniques, sequence mining algorithms, transition graphs or hidden Markov models [13, 19]. All these methods are event-based approaches; where an event is defined as an action of the learner with the course content, tools or learning platform functionalities. However, recently researchers from the Process Mining (PM) field, who are experts in the analysis of data processes, proposed novel methods to unveil learning strategies from big data looking for other representations to understand how self-regulated learning processes occurs [2, 3, 15]. Process Mining techniques can be used to discover models that describe and represent sequences of interactions between learners and course materials [3]. In these recent studies, PM techniques have shown to be very robust to understand users' interactive workflows within a particular system in both structured and unstructured processes. Moreover, compared with other techniques such as sequence mining, transition graphs or hidden Markov models, whose outputs are difficult to relate with natural learning processes and to draw meaningful insights about them. In this sense, PM provides encouraging results for understanding learning processes [6]. Moreover, is a suitable approach for studying learning strategies, as a dynamic regulatory activity carried out during a learning task [25], facilitating the discovery of end-to-end learning process models using the recorded events. But, despite the encouraging results obtained using PM techniques, results from one study do not necessarily apply to other contexts. So, there has been an increasing interest in LA research in replicating studies across contexts [9, 10, 16], although studies of this nature are still scarce in part due to the variation of the instructional conditions [11]. Therefore, new analyses with different data should be done to understand the validity of PM methods in other learning environments and contribute providing more evidence about the impact of the learning context on learners' behaviour and study strategies. To continue this trend of reproducible science, this work builds upon the analytical methodology proposed in a previous study by [18] for unveiling students' learning strategies in self-paced MOOCs in Coursera. In that research, seven different learning strategies were identified, and learners were classified into three groups: samplers, comprehensive, and targeting learners. In the present study, we adapt this particular PM methodology and analyse its application in a MOOC deployed over the edX platform, delivered in a synchronous mode, where the digital resources were developed in English language and consisted in video-lectures, graded and non-graded assessments and other resources. The aim of this adaptation effort is two-fold: (1) to understand whether we could replicate (partially or totally) the analysis conducted in [18] and what methodological decisions we had to change for this purpose and; (2), to extend the current knowledge about students' learning strategies in MOOCs and the influence of the learning context.

2 Related Work

2.1 Analysis of Learning Strategies in MOOCs: Methods and Techniques

To study SRL strategies in online environments, researchers have followed two different approaches: Aptitude-based and Event-based approaches [7]. Aptitude-based approaches offer insights about how learners believe they are using their learning strategies while studying (e.g., self-reports). Event-based approaches conceive learning strategies as a set of events (i.e., actions) that learners perform while they are studying. Event-based approaches overcome some of the weaknesses of aptitude-based approaches, since the former use detailed records of each learner's behaviour, engagement, and other types of interactions with course contents to extract conclusions about their behaviour. However, observing learning strategies in MOOCs, even when these manifest as a set of events or actions, involves several challenges, such as: (a) how to transform traces of fine/coarse-grained data into interpretable behaviour (learning strategies); (b) how to identify and observe behavioural changes; and (c) how to understand whether an observable behaviour relates to a particular learning strategy or to more than one [23].

Recent advances in the evolving disciplines of LA and PM have contributed to overcome these challenges. LA focuses on the human interpretation of data and could provide insights into learning strategies [4], while PM focuses on the application of computational techniques on event-based learning activities to discover sequence of learning behaviour [3]. Examples of these advances are the work done by [21], who applied PM techniques in a MOOC of Coursera with 43,218 learners to understand their learning processes analysing how they performed watching video-lectures and taking assessments. In [18] they used the fuzzy miner algorithm to extract seven types of learning strategies from learners enrolled in four MOOCs of Coursera. Other authors such as [14] used PM to explore learners' quiz-taking behaviour and interaction patterns in a learning management system. Finally, authors in [3] also used PM and clustering techniques to describe the learning behaviour of four groups of learners.

These prior works set the basis to start considering PM as a suitable technique for analysing sequences of learning behaviour. However, more examples and replication studies are needed since both the methodological decisions involved in the use of PM and the context in which the data is gathered may strongly condition the final results.

2.2 Learning Strategies Across Contexts

One of the most important concerns in today's scientific community is that of reproducibility. A key domain in which reproducibility has been identified as a particularly important problem is that of Psychology [23]. Psychology researchers have observed a systematic trend wherein results from studies carried out in one (original) context do not reliably transfer or generalise to other contexts [23, 26]. Examples of contextual factors and changes include everything from demographic variables of participants to the physical or virtual environment in which the study is carried out. This trend has highlighted that fact that results from scientific experiments should always be: (1) sufficiently contextualised and reported on accordingly and (2) replicated across different contexts.

Research in education has found that, just as is the case in Psychology research, the outcomes regarding the impact on learning are also highly dependent on context. Several studies have found that learning outcomes and learner engagement are highly dependent on the context in which the learning occurs [20, 27]. This issue has recently begun to be explored in the LA literature by examining the effect of a course structure/design on passing rates [5]. By leveraging the literature on learning design (the science of structuring and sequencing instructional activities) [17] found that certain course designs (context) lead to significantly different passing rates than others [5]. [8] also demonstrated in a replication study that classifications of learners according to their behaviour varies from a MOOC deployed in Coursera or in FutureLearn, a platform created for promoting a socio-constructivist learning approach [4].

2.3 Research Questions

Two research questions drive this study with the aim of understanding how the methodology for detecting learning strategies proposed in [18] adapts to other learning contexts:

> *RQ1: To what extend can we replicate (partially or totally) the methodology applied in the previous study by* [18] *to extract students' learning strategies in a MOOC?*
>
> *RQ2: How do students' learning strategies in this new context differ from those from the previous study?*

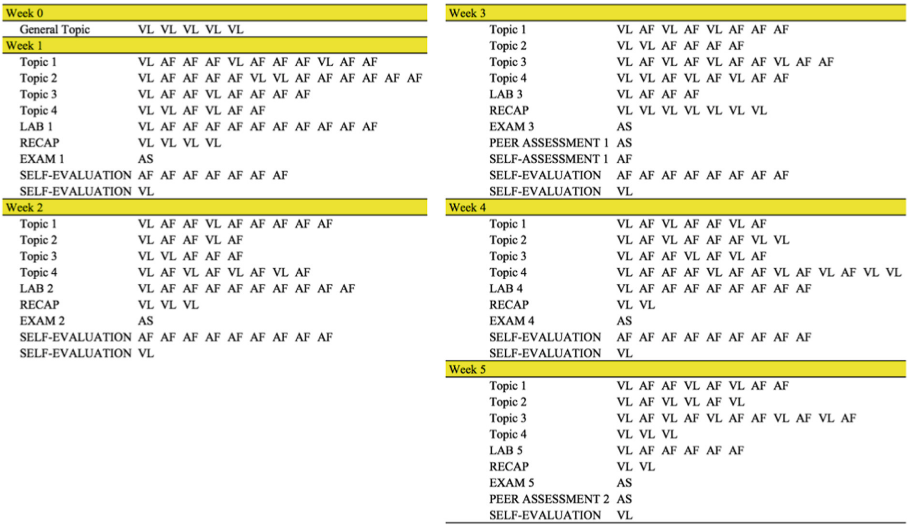
The objective of the *RQ1* is to analyse and discuss what the methodological decisions are needed for applying the same methodology in a different context and see the implications on the final analysis. Regarding *RQ2*, as shown in prior research, learning is highly dependent on context, and the structure and characteristics of a course can have a direct effect on learners' behaviour. In order to understand whether the learning strategies found in [18] vary in this new context, we will analyse two aspects: (1) the learners' behavioural patterns in a synchronous MOOC in edX; and (2) how learners can be classified according to their behaviour and learning outcomes.

3 Method

Some decisions were taken during the process to adapt the methodology developed by [18] to the new learning context. We specified in the text indicating *[Decision-X]*, where "X" corresponds to the number of the methodological decision taken.

3.1 Context: MOOC and Sample

This study used data from one MOOC on "Introducción a la programación en Java" offered by Universidad Carlos III de Madrid in edX. The course was taught in English and the materials were organised into five modules. This MOOC included video-lectures and numerous interactive activities as formative and summative assessments. Figure 1 presents the course structure. This MOOC followed a synchronous approach, and the contents were released weekly. The course was open from April 28th, 2015 until June 30th of the same year. The estimated learners' workload was between five to seven hours per week. To pass the course the learners needed to obtain 60% of the final grade. Summative assessments (exams) had a weight of 75% of the final grade. The rest, 25% of the grade, was assigned to programming activities that consisted of two peer assessments. The final study sample comprised $N = 50{,}776$ online learners that at least completed one video-lecture in the MOOC. The sample selection differs from the study by [18], study in which the subjects were selected based on if they had answered or not a self-reported SRL survey *[Decision-1]*.



Week 0	
General Topic	VL VL VL VL VL
Week 1	
Topic 1	VL AF AF AF VL AF AF AF VL AF AF
Topic 2	VL AF AF AF AF VL VL AF AF AF AF AF AF
Topic 3	VL AF AF VL AF AF AF AF
Topic 4	VL VL AF VL AF AF
LAB 1	VL AF AF AF AF AF AF AF AF AF AF
RECAP	VL VL VL VL
EXAM 1	AS
SELF-EVALUATION	AF AF AF AF AF AF AF
SELF-EVALUATION	VL
Week 2	
Topic 1	VL AF AF VL AF AF AF AF AF
Topic 2	VL AF AF VL AF
Topic 3	VL VL AF AF AF
Topic 4	VL AF VL AF VL AF VL AF
LAB 2	VL AF AF AF AF AF AF AF AF AF
RECAP	VL VL VL
EXAM 2	AS
SELF-EVALUATION	AF AF AF AF AF AF AF AF AF
SELF-EVALUATION	VL

Week 3	
Topic 1	VL AF VL AF VL AF AF AF
Topic 2	VL VL AF AF AF AF
Topic 3	VL AF VL AF VL AF AF VL AF AF
Topic 4	VL VL AF VL AF VL AF AF
LAB 3	VL AF AF AF
RECAP	VL VL VL VL VL VL VL
EXAM 3	AS
PEER ASSESSMENT 1	AS
SELF-ASSESSMENT 1	AF
SELF-EVALUATION	AF AF AF AF AF AF AF AF
SELF-EVALUATION	VL
Week 4	
Topic 1	VL AF VL AF AF VL AF
Topic 2	VL AF VL AF AF AF VL VL
Topic 3	VL AF AF VL AF VL AF
Topic 4	VL AF AF AF VL AF AF VL AF VL AF VL VL
LAB 4	VL AF AF AF AF AF AF AF AF
RECAP	VL VL
EXAM 4	AS
SELF-EVALUATION	AF AF AF AF AF AF AF AF AF
SELF-EVALUATION	VL
Week 5	
Topic 1	VL AF AF VL AF VL AF AF
Topic 2	VL AF VL VL AF VL
Topic 3	VL AF VL AF VL AF AF VL AF VL AF
Topic 4	VL VL VL
LAB 5	VL AF AF AF AF AF
RECAP	VL VL
EXAM 5	AS
PEER ASSESSMENT 2	AS
SELF-EVALUATION	VL

Fig. 1. Structure of the course presenting the contents of each week. VL = video-lecture, AF = formative-assessment, AS = summative-assessment.

3.2 Procedure

To extract students' learning strategies, we followed the stages proposed in [18]. Specifically, they adapted the PM2 methodology [6], and defined four phases to obtain the process model from learners' behaviour in interaction with the course content: (1) extraction stage, (2) event log generation, (3) model discovery and (4) model analysis.

Extraction Stage. The data used in this study were related to learners' commitment with the MOOC contents. These contents were presented in the course as a sequence of different digital resources such as video-lectures, and formative/summative activities. In [18] they only considered interactions with video-lectures and summative activities. In the present study, we extended the data employed to characterise the learners' interaction with the course content by considering the following resources: *LTI* activities (integrating an external development environment called *Codeboard*), graded activities, navigation between modules, tabs and clicks on the home page in edX *[Decision-2]*. Each time a learner interacted with a digital resource in edX, a log with a learning event was generated and stored. This raw data was organised in different files classified in general data, forums, and personal data containing information about learners' behaviour.

Event Log Generation Stage. For creating the event log in this stage, we built upon the two conceptual assumptions defined in [18]: (1) to adopt the same definition of study session as a period of time in which the MOOC platform registered continuous activity of a learner within the course, with intervals of inactivity no greater than 45 m and; (2) to adopt the same definition of an interaction as an event triggered by a learner when this interact with resources from the MOOC. In comparison with the authors in [18], where they defined only six possible interactions, we defined ten types of possible interactions (Table 1) depending on the MOOC structure and the digital resource the learner interacted with *[Decision-3]*. This extension on the number of interactions was a necessary step in order to consider the content provided in the course. Table 1 presents the ten types of interactions defined, which are related to video-lectures, assessments, home view page, and navigation between modules and tabs.

As a result, we defined an event log that contained: (a) the user identification, (b) a time stamp, (c) the interaction performed, and (d) the number of the session in which the event was triggered when learners engaged with MOOC contents. Table 2 presents part of the event log used as an example. We also defined success in a synchronous MOOC based on the grades that learners achieved during the course (at least 60% of the grade in the course), as authors in [18] also did. On the contrary, we did not include the SRL profile as part of the event log *[Decision-4]*.

Model Discovery Stage. Given the exploratory context of this study in which it was necessary to handle complex processes, we selected the same Disco algorithm and their implementation in the Disco commercial tool [12] as authors in [18] also did. The resulting process model was confirmed using the implementation of the Celonis algorithm. Both implementations use a variation in the fuzzy miner algorithm that produced interesting synopses of the learning process in comparison with other techniques [24].

Model Analysis Stage. As a result of the previous stages, we generated a process model that contained learners' behaviour (see Fig. 2). Then, we analysed the observed behaviour in order to unveil learning strategies. For this stage, we identified the most frequent interaction sequences performed by learners that characterised each session, that is the learner' s path followed in the MOOC within a session (see Fig. 3). As authors in [18] did, we ordered the different variants of the sessions from the most common to the least common. The most common ones were assigned to a category that described a session pattern. For example, we analysed the first variants of these sessions and observed that comprised interactions consisting in beginning a video-lecture, then completing or reviewing a video lecture and then ending the session. Therefore, a pattern of "*Only video-lecture*" was defined (i.e., learners working in sessions only with video-lectures).

Table 1. Types of interactions defined based on course resources

Course resource	Interaction	Description
Video-lecture	Begin	Begin but not complete watching a video-lecture that was not previously completed
	Complete	Complete watching more than the 75% of the video-lecture for the first time
	Review	Watch (part of) a video-lecture that was completely watched in the past
LTI activity	Assessment Formative	Attempt to solve a non-graded activity at the first time
	Assessment Formative Review	Go back to a non- graded assessment that was previously visited
Graded activity	Assessment Summative Try	Attempt to solve a graded activity without achieve it
	Assessment Summative Complete	Successful attempt to solve a graded assessment for the first time
	Assessment Summative Review	Go back to a graded assessment that was previously completed successfully
Home Page	Home View	Go to the home page of the course
Modules, Tabs	Navigation	Go through modules (vertically) or tabs (horizontally) looking for specific content

Table 2. Example of the minimal columns of the event log generated

UserId	Time stamp	Interaction	# Session
28	1434522567	Assessment-Formative	1
28	1434522567	Video-Lecture-Complete	1
161	1430520885	Assessment-Formative	1
161	1430520885	Navigation	1
161	1430520885	Navigation	1

Authors in [18] recommend repeating this procedure several times for analysing the rest of the variants in the sessions. This was done using the same Python script developed ad hoc to do this classification task. As a result, we obtained twelve types of sessions (interaction patterns) that learners made.

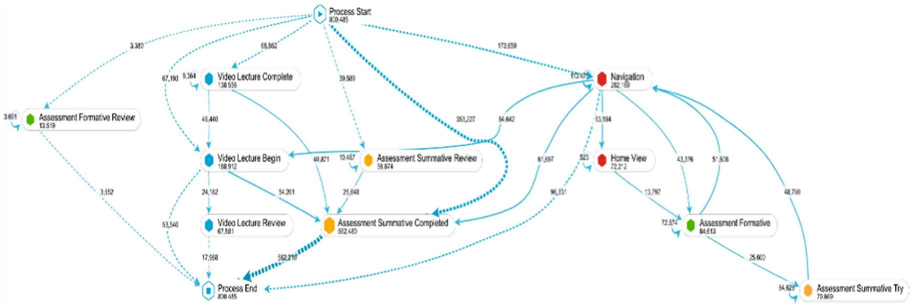
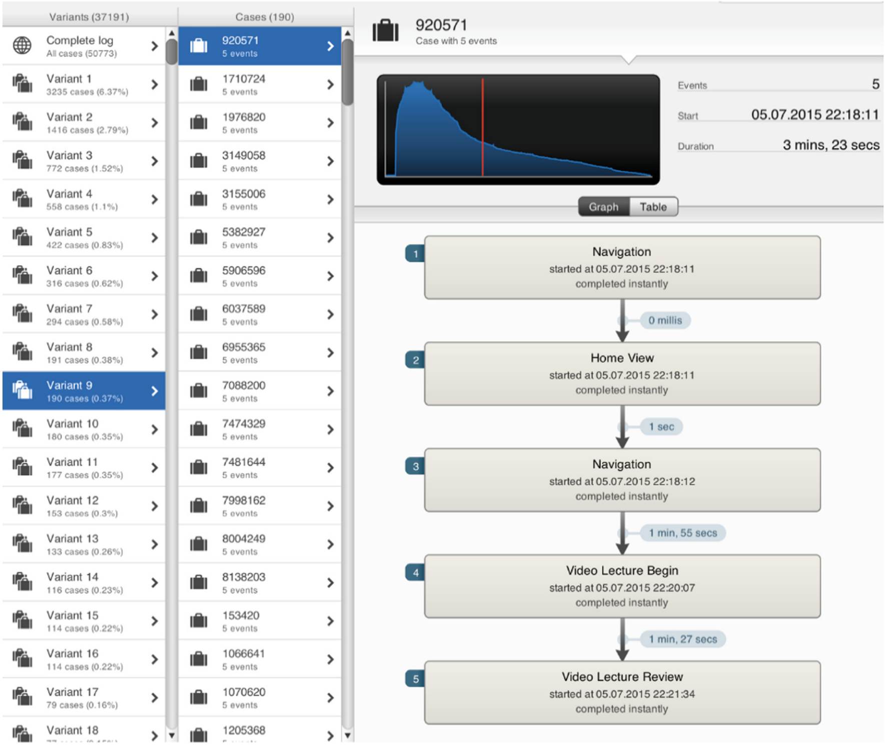


Fig. 2. Full process model obtained using Celonis software, containing all the interactions by sessions. The process model shows ten possible interactions that learner can perform with the course content. Thick dotted line represents the most common path followed by learners.



Variants (37191)	Cases (190)
Complete log All cases (50773)	920571 5 events
Variant 1 3235 cases (6.37%)	1710724 5 events
Variant 2 1416 cases (2.79%)	1976820 5 events
Variant 3 772 cases (1.52%)	3149058 5 events
Variant 4 558 cases (1.1%)	3155006 5 events
Variant 5 422 cases (0.83%)	5382927 5 events
Variant 6 316 cases (0.62%)	5906596 5 events
Variant 7 294 cases (0.58%)	6037589 5 events
Variant 8 191 cases (0.38%)	6955365 5 events
Variant 9 190 cases (0.37%)	7088200 5 events
Variant 10 180 cases (0.35%)	7474329 5 events
Variant 11 177 cases (0.35%)	7481644 5 events
Variant 12 153 cases (0.3%)	7998162 5 events
Variant 13 133 cases (0.26%)	8004249 5 events
Variant 14 116 cases (0.23%)	8138203 5 events
Variant 15 114 cases (0.22%)	153420 5 events
Variant 16 114 cases (0.22%)	1066641 5 events
Variant 17 79 cases (0.16%)	1070620 5 events
Variant 18	1205368

920571
Case with 5 events

Events	5
Start	05.07.2015 22:18:11
Duration	3 mins, 23 secs

Graph Table

1 Navigation
started at 05.07.2015 22:18:11
completed instantly

0 millis

2 Home View
started at 05.07.2015 22:18:11
completed instantly

1 sec

3 Navigation
started at 05.07.2015 22:18:12
completed instantly

1 min, 55 secs

4 Video Lecture Begin
started at 05.07.2015 22:20:07
completed instantly

1 min, 27 secs

5 Video Lecture Review
started at 05.07.2015 22:21:34
completed instantly

Fig. 3. List of the 37,191 variants of sessions obtained using Disco software performed by 50,776 learners in the MOOC. The "variant 9" shows five interactions (events) with four interaction sequences and time associated with the duration of the session.

4 Results

4.1 RQ1: To What Extend Can We Replicate (Partially or Totally) the Methodology Applied in the Previous Study by [18] to Extract Learners' Learning Strategies in a MOOC?

Most of the process in [18] could be applied to the new MOOC. However, some methodological decisions were made to adapt to the structure and data collected in the edX platform, especially in the data-set extraction and log-data construction. In this section we present what were these decisions.

Study Sample [Decision 1]. The study sample of the synchronous MOOC deployed in edX was composed of online learners that at least completed one video-lecture, unlike in the case of the previous study in which the sample was composed of learners who completed an SRL survey. This decision was made because two other previous studies [18, 19] observed that learners' behaviour in the platform was not related with the self-regulatory profile reported in that questionnaire, which is also related to the discussion about the validity of self-reported data in psychological studies [28].

Mapping the Nature of Interactions with Course Resources [Decision 1 and 3]. The MOOC structure of the edX course contained more digital resources compared with the ones in Coursera due to the course design characteristics (video-lectures, formative activities, graded activities, navigation between modules, tabs and clicks on the home page). Accordingly, we mapped the course resources with the possible interactions of the learners and defined ten types of interactions instead of the six defined in the previous study (asynchronous MOOC in Coursera).

Self-reported Information [Decision 4]. This study did not include a self-reported SRL profile of the students (as it was done in [18]) as part of the event log. This variable was found to not have an influence in the process of exploring the patterns of the behaviour found. However, knowing the self-reported profile of the learners helps to have a better understanding of the characteristics of the students and relate their profile to their actions. To sum up, these four decisions lead us to adapt the methodology used in [18] in the context of this study.

4.2 RQ2: How do Learners' Learning Strategies in This New Context Differ from Those from the Previous Study?

To answer this research question, two analyses were conducted. Next, we present the results of these analyses.

a) Analysis of Learners' Behavioural Patterns in a Synchronous MOOC in edX. We obtained twelve types of interaction sequence patterns that learners made when they engaged with the MOOC (see Table 3). The description of each interaction sequence pattern was grounded upon whether a session only contained a certain type of interaction (e.g., sessions consisting of *only-video-lectures* without any assessment activity) or whether the session contained certain type of interaction sequences between interactions that are considered important for the learning process (e.g., sessions where learners went from *trying a summative-assessment* to a *video-lecture* activity). Once the most common sessions patterns were extracted from the main process model (see Fig. 2), we obtained a specific process model for each pattern (see example in Fig. 4). Twelve distinct types (patterns) of sessions were extracted: ***(1) Only assessment-summative-complete***: Session pattern in which learners worked only passing graded assessments. This is the most common type of session: 44.11% of the total number of sessions corresponded to this type. ***(2) Only video-lecture to assessment-summative-complete***: Session pattern in which learners began working with video-lectures (either beginning, completing) and then successfully solved a graded assessment (summative) for the first time (see Fig. 4): 13.44% of the sessions corresponded to this type.

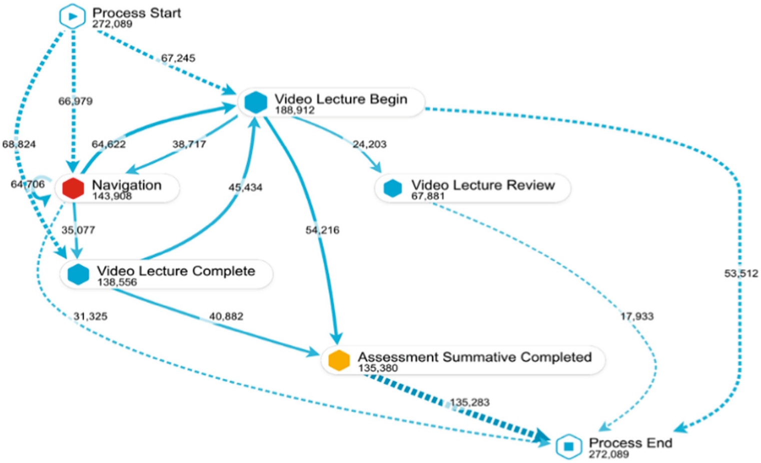


Fig. 4. Only video-lecture to assessment summative complete session pattern performed by learners in a MOOC

(3) Only video-lecture: Session pattern in which learners worked only with video-lectures. Learners performed sessions that consisted of watching at least one video-lecture and did not contain assessment activities. Learners could begin, complete, review video-lectures or perform combinations of them (i.e., begin and then complete, begin and then review, complete and then review): 10.78% of the sessions corresponded to this type. ***(4) Only assessment-summative***: Session pattern in which learners worked only with summative assessments. Learners performed sessions that consisted in trying at least one summative assessment and did not watch any video-lecture. Learners could try, complete, review summative assessments or performed combinations of them (i.e., try and then complete, try and then review, complete and then review) while they were interacting with the course:10.03% of the sessions corresponded to this type. ***(5) Only assessment-formative***: Session pattern in which learners worked only with formative assessments. Learners performed sessions that consisted of attempting at least one formative assessment and did not watch any video-lecture. Learners could attempt or review formative assessments or perform combinations of them (i.e., attempt an assessment and then end the session, attempt and then review, review and then end the session): 9.59% of the sessions corresponded to this type. ***(6)Combined:*** Session pattern in which learners combined from two up to four sessions patterns mentioned in this section: when the combination is up to two, all types of sessions were considered as part of this combined session pattern; when the combination is up to three, sessions consisting in work only with video-lectures and only with assessments were not considered as part of this combined session pattern; when the combination is up to four, sessions consisting in working only with video-lectures, only with assessments and explore were not considered as part of this combined session pattern: 4.15% of the sessions corresponded to this type. ***(7) Only-assessment:*** Session pattern in which learners worked between formative and summative assessments in the same session. Learners could attempt to solve or review a non-graded assessment activity (formative) and try to complete (pass) a graded assessment activity (summative) while they were interacting with the course: 2.27% of the sessions corresponded to this type. ***(8) Only video-lecture to assessment-formative:*** Session pattern in which learners began working with video-lectures (either beginning, completing or reviewing) and then attempted to solve a non-graded activity at the first time: 2.24% of the sessions corresponded to this type. ***(9) Explore:*** Session pattern in which learners worked only beginning video-lectures (without completing) or attempting some non-graded formative assessments. ***(10) Assessment-summative-try to Only-video-lecture:*** Session pattern in which learners attempted to solve a graded activity incorrectly and then worked with video-lectures (begin, complete, review video-lectures or combinations of them). ***(11) Video-lecture-complete to assessment-summative-try:*** Session pattern in which learners completed a video-lecture and then attempted to solve a graded activity without managing to do it. ***(12) Others:*** We have classified as other to those sessions that were long and disperse, as they do not fit into any of the above-mentioned session patterns.

Table 3. Percentage of session patterns performed by learners (N = 800,485 sessions)

Session patterns	# Sessions (%)
(1) Only assessment-summative-complete	353,090(44.11%)
(2) Only video-lecture → assessment-summative-complete	107,623(13.44%)
(3) Only video-lecture	86,306(10.78%)
(4) Only assessment-summative	80,310(10.03%)
(5) Only assessment-formative	76,791(9.59%)
(6) Combined	33,253(4.15%)
(7) Only assessment	18,205(2.27%)
(8) Only-video-lecture → assessment-formative	18,000(2.24%)
(9) Explore	10,095(1.26%)
(10) Assessment-summative-try → only-video-lecture	9,463(1.18%)
(11) Others	6,644(0.83%)
(12) Video-lecture-complete → assessment-summative-try	705(0.08%)

b) Learners' Classification According to Their Behaviour and Learning Outcomes. To answer this question learners ($N = 50{,}776$) were grouped based on the identified sessions patterns. We use the agglomerative hierarchical clustering as in [18]. The resulting dendrogram was used to identify the optimal number of clusters (qualitative). Then, using the *Gaussian mixture* and *K-means* clustering techniques, we confirmed the number of clusters based on the silhouette score (quantitative). This led to selecting the solution with four clusters (see Fig. 5). Table 4 describes the resulting clusters in terms of: (a) the ten session patterns used for grouping the learners (we discarded video-lecture-complete to assessment-summative-try and others given that both types are less than 1% of all sessions), (b) the mean in terms of session performed, (c) the number of learners, (d) the number of learners that passed/failed the course.

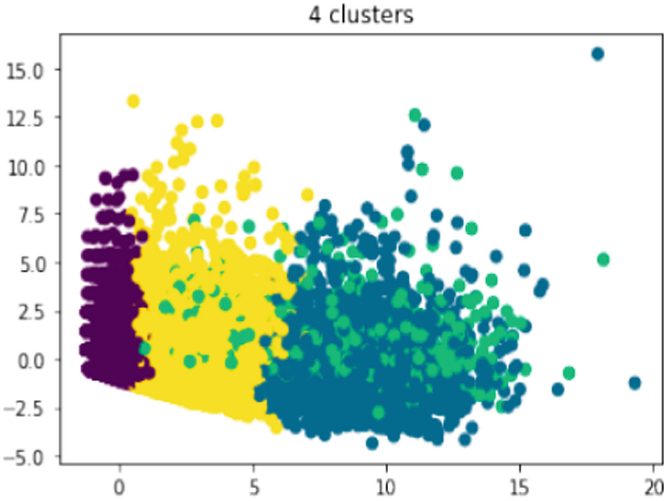


Fig. 5. Scatter Plot with silhouette score 0.571

The resulting clusters indicate different types of learning strategies deployed by learners while they were facing the MOOC. If we look for specific differences between the different clusters, we can describe them as follows (see Table 4; Table 5; Table 6 and Fig. 5):

Cluster 0 – Sampling Learners: This cluster was composed of learners that on average visited only once or twice the course exploring the course content. Specifically, they visited the video-lectures and follow through the proposed path by the course to visit formative assessments but without attempting or ending any activity proposed, just exploring the content to see the big headlines. This cluster is composed of the largest number of learners ($n = 30{,}415$), but they fail passing the course.

Cluster 1 – Targeting Learners: This cluster was composed of learners that on average performed a low number of sessions. Although they were active learners, they had low activity in the course in comparison with the next groups (clusters 2 and 3, see Table

Table 4. Means of session patterns per cluster performed by learners (N = 800,485)

Session patterns	Cluster 0	Cluster 1	Cluster 2	Cluster 3
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Assessment-summative-try → only-video-lecture	.00(.07)	.29(.62)	1.49(1.63)	1.62(1.68)
Combined	.01(.11)	1.26(1.33)	3.91(3.40)	4.08(3.36)
Explore	.14(.38)	.23(.56)	.39(.66)	.62(.85)
Only-assessment-summative-complete	.00(.00)	**10.79**(13.66)	**47.03**(32.67)	**69.15** (27.79)
Only-assessment	.00(.08)	.74(1.10)	1.43(1.31)	2.04(1.35)
Only-assessment-formative	**.76**(.095)	1.78(2.31)	5.70(4.90)	**9.54**(5.42)
Only-assessment-summative	.00(.07)	**2.65**(3.88)	**9.96**(6.93)	**14.03**(6.19)
Only-video-lecture	**.33**(.68)	**2.22**(3.18)	**9.94**(8.63)	**15.90**(9.63)
Only-video-lecture → assessment-formative	**.37**(.617)	.32(.73)	.40(.72)	.37(.64)
Only-video-lecture → assessment-summative-complete	.00(.00)	**2.86**(4.30)	**15.92**(11.81)	**24.59**(10.64)
N_sessions_on_average_per_cluster	1.69(1.37)	23.37(25.59)	97.77(61.68)	142.761(45.77)
N_learners	**30,415**	**17,829**	**651**	**1,881**
Fail_course	30,415	17,786	492	1,005
Pass_course	**0**	**43**	**159**	**876**

4, Table 5 and Table 6). They worked superficially with the course materials. These learners after watching video-lectures attempted to pass summative assessments leaving formative assessment aside (sessions were mainly oriented to passing the summative assessments). This behaviour shows that learners in this cluster focused on passing the course more than on achieving a deep understanding of the contents and self-evaluating their progress. This cluster is composed of a great number of learners (n = 17,829), but only a few of them passed the course (n = 43, compared with clusters 2, 3).

Cluster 2 – Low Comprehensive Learners: This cluster was composed of learners that on average performed a large number of sessions in comparison with the previous two groups (clusters 0, 1). They worked intensively with the course materials. These learners watched the video-lectures, attempted formative and then summative assessments (which is the path designed by the instructors in the course). They focused on summative more than formative assessments (see Table 4, Table 5 and Fig. 6). Also, after watching video-lectures they intended to pass summative assessments and worked less with formative assessments (in comparison with cluster 3). However, learners in this cluster performed more sessions working with summative assessments than with formative ones. In this cluster, a large number of learners passed the course (n = 159, in comparison with cluster 1).

Cluster 3 – Highly Comprehensive Learners. This cluster was composed of learners that on average performed a large number of sessions and worked with more intensity with the course contents than learners in the rest of the clusters (see Table 4; Table 5 and Fig. 6). Learners in cluster 3 performed more sessions that consisted in working with video-lectures before they passed a summative assessment. Also, they performed more sessions either with formative or summative assessments in comparison with learners in cluster 2. This behaviour showed the intention of learners to achieve a deep understanding of the contents and self-evaluate their progress. Learners in this cluster also performed sessions in which they worked intensively only with video-lectures in comparison with the rest of the learners in the different clusters.

Finally, Table 6 presents comparisons between the four clusters based on the distributions of the session patterns. Between clusters 2 and 3 there are no statistically significant differences, while pair comparisons between clusters 0–1, 1–2, 1–3 showed statistically significant differences.

To analyse the relationship between students' learning behavioural patterns and their performance in the course we followed the same methodology applied in [18] and compared how learners performed the different sessions patterns depending on their achievements (passing or not passing the course). However, learners in clusters 2 and 3, classified as low and highly comprehensive learners respectively, behaved differently in terms of passing the course. Although learners in these clusters worked on average the same number of sessions in the course (no statistical differences observed), their study strategies differ (Table 7).

Table 5. Differences in session patterns between cluster 2–3

Session patterns	Cluster 2	Cluster 3	*t*	*p*	*r*
	M	*M*			
Assessment-summative-try → only-video-lecture	1.49	1.62	−.95	.342	.06
Combined	3.91	4.08	−1.44	−.49	.07
Explore	.39	.62	−6.94	**< .001*** **	.39
Only-assessment-summative-complete	**47.03**	**69.15**	−8.02	**< .001*** **	.49
Only-assessment	1.43	2.04	−5.38	**< .001*** **	.33
Only assessment-formative	5.70	**9.54**	−8.91	**< .001*** **	.50
Only assessment-summative	**9.96**	**14.03**	−6.91	**< .001*** **	.43
Only video-lecture	**9.94**	**15.90**	−7.86	**< .001*** **	.45
Only-video-lecture → assessment-formative	.40	.37	.50	.61	.03
Only video-lecture → assessment-summative complete	**15.92**	**24.59**	−8.63	**< .001*** **	.51
N_sessions_on_average_per_cluster	97.77	142.76	−17.05	**< .001*** **	.49
N_learners	**651**	**1,881**			

Note: *** p < *.001;* marks statistically significant differences.

Table 6. Comparison between clusters of learners based on the session patterns

Cluster #	Cluster #	χ2	*p*
0	1	281.35	< .001***
1	2	194.99	< .001***
1	3	529.99	< .001***
2	3	15.18	.231

Note: *** p < *.001* marks statistically significant differences.

Highly comprehensive learners (cluster 3): (a) worked more in sessions that consisted in watching video-lectures and then passing summative assessments, (b) worked more with formative assessments and worked in combination with summative and formative assessments, and (c) on average explored more the course contents.

In contrast, *low comprehensive learners* (cluster 2): (a) worked more in sessions in which they tried to pass a summative assessment (but failed) and then went back to work with video-lectures (begin, complete or review), and (b) worked more with combinations of the different session patterns in comparison with highly comprehensive learners. In addition, *low comprehensive learners* tried to pass summative assessments but when failing, they work in video-lectures, probably trying to find information in the video-lectures that helped them to pass the summative assessments. In contrast, *highly comprehensive learners* worked first with video-lectures and then passed summative assessments. This behaviour suggests that this type of learner is trying to achieve a deep understanding of the contents and self-evaluate their progress working more with formative assessments.

Table 7. Differences in session patterns performed on average by learners in clusters 2–3 that passed the course

Session patterns	Cluster 2 (pass)	Cluster 3 (pass)	*t*	*p*	*r*
	M	*M*			
Assessment-summative-try → only-video-lecture	2.25	1.49	4.89	**< .001*** **	.32
Combined	5.89	3.82	6.24	**< .001*** **	.40
Explore	.34	.50	−3.07	**< .002** **	.17
Only-assessment	1.90	2.12	−2.04	**< .045** **	.13
Only-assessment-formative	10.94	11.85	−2.08	**< .038** **	.14
Only-video-lecture → assessment-summative complete	30.62	31.85	−2.20	**< .028** **	.13
N_sessions_on_average_per_cluster	88.81	88.814	−.0029	.998	.000
N_learners	**159**	**876**			

Note. ** $p < .05$, *** $p < .001$ marks statistically significant differences.

5 Conclusion and Discussion

5.1 Summary of Results

Even if conducting the same study across different context is complicated by variations in instructional conditions [11], in this study we made an effort of replicability and applied the PM methodology in [18] to a data set of a synchronous MOOC in the edX platform. Two main results were obtained. Firstly, the PM methodological approach can be replicated, but it requires taking three key decisions that are dependent to the context of application: (1) the sample size, which will vary from experiment to experiment; (2) mapping the nature of the interactions based on the structure of the MOOC under analysis, but keeping the metric of session and interaction; and (3) eliminating students' SRL profile obtained from a SRL-questionnaire as a control measure. Secondly, the adaptation of this methodological approach extends the findings in [18] by identifying new learning strategies that are highly dependent on the course structure. In contrast to the six self-regulatory patterns and three groups of learners identified in the prior work, we identified twelve patterns and four groups: 1) Sampling learners, 2) Targeting learners, 3) Low Comprehensive learners, and 4) Highly Comprehensive learners.

5.2 Implications

The present findings have implications both for (a) the methods used in the LA community for analysing trace data, and (b) for theory and practice of SRL. **Regarding the implications in LA methods**: This paper sheds some light on the aspects to be considered when doing replication studies using students' trace data. Replicating an analytical method requires taking decision about how raw data is processed. In order to evaluate the reproducibility of the results, these decisions should be carefully reported, especially when they require some level of pre-processing or abstraction. When applying PM approaches, the data pre-processing and data abstraction is key. For example, how students' work session is defined or how student's interactions with the course content are mapped into a logfile may have an impact on how learners' strategic patterns are observed. This study shows that, when replicating methodological approaches based on PM, the granularity of the data when defining students' interaction should maintained from one study to another. That is, if student's interaction with the course content is defined by interaction with a particular resource, this should be the level of granularity for the analysis, and no combinations of interactions should be considered for the analysis. In current literature, most of studies take as a reference the interactions with the course content as a basis [13, 24], however, this could vary when changing platform, since the nature of the data collected may vary. The results of this study emphasize the importance of including the decision-making process on data preprocessing as part of any analysis in order to be able to compare the results from one study to another. Moreover, this pre-processing should consider simplifying the raw data by keeping only those types of interaction that could be translated from one platform to another, even if this means losing some data in the process. Of course, simplifying the data may mean also simplifying the results, but more studies of this type should be reported so that the community arrives to agreements such as a standard of a minimum logfile to facilitate replication studies. **Regarding the implications for SRL theory and practice:** The adaptation of this methodology extends the findings in [18] by identifying new learning strategies that are highly dependent on the course structure. Twelve sessions patterns and four groups of learners were found. Learners classified as sampling and targeting in this study are similar to those found in [18]. However, in contrast to the prior work, *Comprehensive learners* can be classified into *highly* and *low* comprehensive. *Highly comprehensive learners* seemed to be deeper learners following the designed path of the course, trying to achieve a deep understanding of the contents and self-evaluating their progress through the intensive work with formative activities. In contrast, *low comprehensive learners* seemed to be more strategic, following a pattern that consisted in passing summative activities and working less with formative ones. While in the prior work, [18] analysed a MOOC with only summative assessment activities, the MOOC in the present study included more than 160 formative activities. These results suggest that the strategies adopted by the learners are highly dependent on the context, and in particular, on the course content and structure. Moreover, these results align with prior work that show how course structure and design conditions students' behaviour [1, 17]. However, more studies, and particular A/B experimental experiments, should be conducted in order to provide robust evidences on how context affects learners' behaviour. Moreover, and beyond replication efforts, we believe that the identified behavioural patterns can inform the design of learning environments by either supporting the implementation of precise learner modelling or by providing enough scaffolding to at-risk learners who remain working actively in the MOOC.

5.3 Limitations

The findings of this study are subject to some limitations given the nature of the data and methodological choices. First, this study is based on learners' behavioural data that were automatically collected by the MOOC platform, so the analyses are limited to the data provided. Second, and for the effort of replication, the set of interactions obtained for analysing the learners' behaviour through study sessions were simplified to consider video-lectures, assessments (either summative or formative) and navigation interactions. Considering data such as the students' forum activity may alter the strategic patterns encountered. Future work will expand this study considering data from collected by other researchers from other courses and platforms in order to conduct a meta-analysis following the same methodology.

Acknowledgments. This paper was supported by the ANR LASER (156322) and Vicerrectorado de Investigación de la Universidad de Cuenca. The authors acknowledge PROF-XXI, which is an Erasmus+ Capacity Building in the Field of Higher Education project funded by the European Commission (609767-EPP-1-2019-1-ES-EPPKA2-CBHE-JP).

References

1. Alario-Hoyos, C., et al.: Understanding learners' motivation and learning strategies in MOOCs. Int. Rev. Res. Open Distrib. Learn. **18**(3), 119–137 (2017)
2. Alonso-Mencía, M.E., et al.: Self-regulated learning in MOOCs: lessons learned from a literature review. Educ. Rev., 1–27 (2019)
3. van den Beemt, A., et al.: Analysing structured learning behaviour in massive open online courses (MOOCs): an approach based on process mining and clustering. Int. Rev. Res. Open Distrib. Learn. **19**(5) (2018)
4. Boekaerts, M.: Self-regulated learning: a new concept embraced by researchers, policy makers, educators, teachers, and students. Learn. Instr. **7**(2), 161–186 (1997)
5. Davis, D., et al.: Toward large-scale learning design: categorizing course designs in service of supporting learning outcomes. In: Proceedings of the Fifth Annual ACM Conference on Learning at Scale, p. 4 (2018)
6. van Eck, M.L., Lu, X., Leemans, S.J.J., van der Aalst, W.M.P.: PM^2: A process mining project methodology. In: Zdravkovic, J., Kirikova, M., Johannesson, P. (eds.) CAiSE 2015. LNCS, vol. 9097, pp. 297–313. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19069-3_19
7. Endedijk, M.D., Brekelmans, M., Sleegers, P., Vermunt, J.D.: Measuring students' self-regulated learning in professional education: bridging the gap between event and aptitude measurements. Qual. Quant. **50**(5), 2141–2164 (2015). https://doi.org/10.1007/s11135-015-0255-4
8. Ferguson, R., Clow, D., Beale, R., Cooper, A.J., Morris, N., Bayne, S., Woodgate, A.: Moving through MOOCS: pedagogy, learning design and patterns of engagement. In: Conole, G., Klobučar, T., Rensing, C., Konert, J., Lavoué, É. (eds.) EC-TEL 2015. LNCS, vol. 9307, pp. 70–84. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24258-3_6
9. Ferguson, R., Clow, D.: Consistent commitment: patterns of engagement across time in massive open online courses (MOOCs). **2**, 55–80 (2015). https://doi.org/10.18608/jla.2015.23.5
10. Gardner, J., et al.: Replicating MOOC predictive models at scale. In: L@S, p. 1 (2018)
11. Gašević, D., et al.: Learning analytics should not promote one size fits all: the effects of instructional conditions in predicting academic success. Internet High. Educ. **28**, 68–84 (2016)
12. Günther, C.W., Rozinat, A.: Disco: discover your processes. BPM (Demos). **940**, 40–44 (2012)
13. Jovanović, J., et al.: Learning analytics to unveil learning strategies in a flipped classroom. Internet High. Educ. **33**, 74–85 (2017)

14. Juhanák, L., et al.: Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. Comput. Hum. Behav. **92**, 496–506 (2017)
15. Kizilcec, R., Pérez-Sanagustín, M., Maldonado, J.J.: Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. Comput. Educ. **104**, 18–33 (2017). https://doi.org/10.1016/j.compedu.2016.10.001
16. Kizilcec, R.F., Brooks, C.: Diverse big data and randomized field experiments in MOOCs. Handb. Learn. Analytics, 211–222 (2016)
17. Laurillard, D.: Teaching as a design science: Building pedagogical patterns for learning and technology. Routledge (2013)
18. Maldonado-Mahauad, J., et al.: Mining theory-based patterns from big data: identifying self-regulated learning strategies in massive open online courses. Comput. Hum. Behav. **80**, 179–196 (2018)
19. Matcha, W., et al.: Detection of learning strategies: A comparison of process, sequence and network analytic approaches. In: Scheffel, M., Broisin, J., Pammer-Schindler, V., Ioannou, A., Schneider, J. (eds.) EC-TEL 2019. LNCS, vol. 11722, pp. 525–540. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29736-7_39
20. Meyer, J.H.F., Muller, M.W.: Evaluating the quality of student learning. I—an unfolding analysis of the association between perceptions of learning context and approaches to studying at an individual level. Stud. High. Educ. **15**(2), 131–154 (1990)
21. Mukala, P., et al.: Exploring students' learning behaviour in MOOCs using process mining techniques. Eindhoven. BPM Center Report BPM-15–10. BPMCenter.org, Eindhoven Google Scholar (2015)
22. Pardo, A., et al.: Generating actionable predictive models of academic performance. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK 2016, pp. 474–478 (2016). https://doi.org/10.1145/2883851.2883870
23. Pashler, H., Wagenmakers, E.-J.: Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? Perspect. Psychol. Sci. **7**(6), 528–530 (2012)
24. Saint, J., Gašević, D., Pardo, A.: Detecting learning strategies through process mining. In: Pammer-Schindler, V., Pérez-Sanagustín, M., Drachsler, H., Elferink, R., Scheffel, M. (eds.) EC-TEL 2018. LNCS, vol. 11082, pp. 385–398. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98572-5_29
25. Sonnenberg, C., Bannert, M.: Discovering the effects of metacognitive prompts on the sequential structure of SRL-processes using process mining techniques. J. Learn. Analytics **2**(1), 72–100 (2015)
26. Stanley, D.J., Spence, J.R.: Expectations for replications: are yours realistic? Perspect. Psychol. Sci. **9**(3), 305–318 (2014)
27. Trigwell, K., Prosser, M.: Improving the quality of student learning: the influence of learning context and student approaches to learning on learning outcomes. High Educ. (Dordr) **22**(3), 251–266 (1991)

28. Veletsianos, G., et al.: The life between big data log events: learners' strategies to overcome challenges in MOOCs. AERA Open **2**(3), 2332858416657002 (2016). https://doi.org/10.1177/2332858416657002
29. Winne, P.H., Hadwin, A.F.: Studying as self-regulated learning. Metacognition Educ. Theor. Pract. **93**, 27–30 (1998)