JOURNAL OF LEARNING ANALYTICS
SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH
Volume 10(3), 7–27. https://doi.org/10.18608/jla.2023.7911

Session-Based Time-Window Identification in Virtual Learning Environments

Aleksandra Maslennikova[1], Daniela Rotelli[2], Anna Monreale[3]

Abstract
Students organize and manage their own learning time, choosing when, what, and how to study due to the flexibility of online learning. Each person has unique learning habits that define their behaviours and distinguish them from others. To investigate the temporal behaviour of students in online learning environments, we seek to identify suitable time-windows that could be used to investigate their temporal behaviour. First, we present a novel perspective for identifying different types of sessions based on individual needs. The majority of previous works address this issue by establishing an arbitrary session timeout threshold. In this paper, we propose an algorithm for determining the optimal threshold for a given session. Second, we use data-driven methods to support investigators in determining time-windows based on the identified sessions. To this end, we developed a visual tool that assists data scientists and researchers to determine the optimal settings for session identification and locating suitable time-windows.

Notes for Research

- In this paper we tackle the problem of identifying appropriate time-windows that could be used to investigate the temporal behaviour of students in online learning environments and to better adapt analysis techniques to a given dataset.
- Previous research has often identified time-windows intuitively or based on personal experience and viewpoints. In contrast to previous research, we propose a method to support investigators in identifying time-windows objectively using a data-driven approach based on the concept of session, which we have reformulated in three different forms to meet various individual requirements.
- We also introduce an algorithm for estimating the duration of inactivity, i.e., *off-task* activity, during online learning.
- To identify time-windows, we developed a visual tool, whose whole source code is freely available, to assist data scientists and researchers in determining the optimal settings for the session identification and locating suitable time-windows.

Keywords
Session identification, time-window identification, session timeout threshold, time-on-task, time-off-task.

Submitted: 20/10/2022 — **Accepted:** 20/09/2023 — **Published:** 15/12/2023

[1] *Email: a.maslennikova@studenti.unipi.it Address: Department of Computer Science, University of Pisa, Pisa, Italy. ORCID ID: https://orcid.org/0000-0002-5877-9411*
[2] *Corresponding author Email: daniela.rotelli@phd.unipi.it Address: Department of Computer Science, University of Pisa, Pisa, Italy. ORCID ID: https://orcid.org/0000-0002-0943-6922*
[3] *Email: anna.monreale@unipi.it Address: Department of Computer Science, University of Pisa, Pisa, Italy. ORCID ID: https://orcid.org/0000-0001-8541-0284*

1. Introduction

Study skills, also known as *study behaviour* or *learning strategies*, are defined as the ability to manage time and other resources in line with the demands of academic work and the ability to organize, summarize, and integrate material (Credé & Kuncel, 2008). Learning occurs over time, and time management is a strategic learning component of *self-regulation* (Weinstein et al., 2002), which implies that students develop the ability to self-regulate their study by choosing when, where, what, and how they study (Schmitz & Wiese, 2006).

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

In online learning environments (OLEs), event logs describe interactions that occurred on a platform and provide multiple dimensions that help to characterize what actions students take, when, how long, and where (in which course and in which part of the course). Learning analytics (LA) research has demonstrated that temporal analysis is relevant since the greater timing flexibility in online environments may affect learning differently (Riel et al., 2018). A number of works rely on the analysis of logs to investigate the temporal dimension of learning (Sher et al., 2022; Khan & Pardo, 2016; Chen et al., 2018; Y. Lee, 2018; Rotelli & Monreale, 2022; Wang et al., 2023). Therefore, analyzing student temporal learning behaviour may provide valuable insights to instructional designers in improving a course's design and making it more effective for students. Moreover, the results of those analyses could be helpful for designing dashboards supporting both *teachers* in understanding and monitoring the types of timing strategies and behaviours that learners employ in their self-paced coursework and *students* in monitoring, controlling, and acquiring awareness of their actual habits in relation to their perception.

To investigate the heterogeneous nature of various online temporal study behaviours, we have to define a profile tailored to the temporal analysis. As humans, we operate under the cadence of a seven-day week, and the week-day alternation and the time of day drive our lives. Since the flexibility of e-learning requires student self-regulation (time management and self-testing) (Goda et al., 2015), some questions arise. One of them is how students act at different times of the day. Therefore, we look at identifying time-windows that could help us analyze students' temporal behaviour when learning online.

Nevertheless, when examining our data, we faced the difficulty of defining the most representative time segmentation of student temporal learning habits. One way to approach this problem is to designate time periods *intuitively* (Sher et al., 2022) or depending on personal experience and perspectives (Du et al., 2019). To prevent making arbitrary selections, starting from the definition of a *task*, in this work we propose a data-driven methodology for supporting time-window identification based on (i) the definition and extraction from logs of three distinct *session* types that we identified: *study*, *course*, and *learning sessions*, and (ii) a visual tool that, exploiting these sessions, supports the identification of time-windows tailored to the data being analyzed.

The remainder of this paper is structured as follows. Section 2 provides an overview of prior research on session identification and time segmentation. Section 3 describes the dataset used for the data analysis and details how the data were prepared for analysis. In Section 4 we formulate the problem we want to tackle in this paper. In Section 5, preliminaries that will help the reader comprehend the specifics of our proposal are defined. Section 6 illustrates the approach we propose to use to identify time-windows based on visual analytics of sessions as well as its strengths and limitations. Finally, Section 7 reports a conclusion and research directions for the future.

2. Background and Related Work

Since its creation, the World Wide Web has undergone significant development. It has transformed from a collection of static content to a platform where users can find information, contribute, interact, and collaborate. Technological im provements have made it possible to record users' browsing behaviours through web log files and web usage mining (WUM), defined as "the automatic discovery of user access patterns from Web servers" (Cooley et al., 1997), in order to identify interesting patterns from web usage data (Lokeshkumar et al., 2014). Numerous studies dealing with handling and using interaction logs in research evaluate how users interact with search systems to capture search behaviour (Catledge & Pitkow, 1995).

Because the web log file contains a considerable quantity of unnecessary information, which cannot be used directly in the WUM process, data pre-processing is a fundamental phase (Chitraa & Davamani, 2010). The pre-processing phase includes (i) *data cleaning*, which involves the removal of irrelevant log records; (ii) *user identification*, typically based on the IP addresses and web user agent resolution; (iii) *session identification*, which aims at separating the stream of user log data into individual access sessions, i.e., into separate visits to the site (Spiliopoulou et al., 2003); and (iv) *path completion*, which addresses the problem of missing information in the server access log.

Prior research has proposed various strategies to determine appropriate session boundaries, such as *time-oriented heuristics*, imposing a threshold on the total session duration (usually 30 minutes) or a maximum duration limit on a specific page (generally 10 minutes) (Cooley et al., 1999); *navigation-based heuristics*, making use of browser navigation, i.e., examining the structure of the website to identify user sessions: when two sequential logs refer to pages which are not directly connected, a new user session begins (Nadjarbashi-Noghani & Ghorbani, 2004); *task-based heuristics*, finding groups of semantically related search queries tackling the same task (Boldi et al., 2008; Jones & Klinkner, 2008); or employing meaningful *transactions* identified using a reference length approach (Fatima et al., 2016).

In computer science, a *session* is an information exchange between two or more devices that lasts for a certain amount of time. In library search systems, a session, also called a *login session*, a *user session*, or an *access session*, is the time between a user logging in and logging out (Jones & Klinkner, 2008); therefore sessions are easily identifiable through login IDs. On the Internet, however, there is usually no official login and logout to access and use search engines. In the case of a shared computer, identifiers such as IP addresses and cookies may also be shared by several users. Consequently, the term *session* has

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

been split between multiple meanings: (1) a set of queries designed to serve a specific information demand, (2) a sequence of consecutive questions, and (3) a brief period of time spent querying and examining responses (Jones & Klinkner, 2008). Nevertheless, a session is also a single visit to the site (Asadianfam & Mohammadi, 2014) or the collection of web pages viewed by a single user for a specific purpose (Jafari et al., 2013). A *browser session* is a continuous period of user activity with no more than a 30-minute gap between events (Cooley et al., 1999). For a web search, usually, a session is addressed as specific brief sequences of queries submitted to the web search engine by a user to address a single information need (Jansen et al., 2007), and *search sessions* are temporal sequences of queries issued by users (Lucchese et al., 2013). However, there have been many attempts to segment and define sessions that do not directly address the concept of user information demands but rely on the concept of similar context, topic, or temporal characteristics (Jones & Klinkner, 2008; Fatima et al., 2016).

In the educational setting, a session might assume various facets. A *reference session* indicates a user session (Marquardt et al., 2004). A *learning session* is a user session that is focused on a specific learning activity and can last several days (Marquardt et al., 2004), but it has also been defined as a set of learning interactions that the student engages in to complete a learning task (Heraud et al., 2004). A *study session* is described as a learning process in which the student engages in learning activities consecutively while maintaining the educational context of their prior activities (Y. Lee et al., 2020). However, it has also been defined as the sequence of occurrences that occur within 21 minutes of each other (Fincham et al., 2019); the sequence of a student's access to online materials throughout a different period of the year with a fixed duration (one day, three days, one week, and one month) (Brooks et al., 2015); a period of time during which the trace data capture the continuous activity of a learner within a course, with inactivity intervals of no more than 45 minutes (Maldonado-Mahauad et al., 2018); and the time for the student to go offline for coding or paper problem solving and then to return online (Gitinabard et al., 2019). The activity of a student on the platform has been classified as a *user session* (Kovanović et al., 2015), a *study session* (Khan & Pardo, 2016), a *browser session* (Gitinabard et al., 2019), or a *learning session* (Jovanovic et al., 2017). Based on this outline, there appears to be an evident lack of standardized terminology. However, when online, a student can navigate the platform for a general overview, actively participate in a course, or engage in effective learning. Hence, in our opinion, distinguishing between *study session*, *course session*, and *learning session* would be beneficial, as we will formally define in Section 5.1.

In light of the move to web-based learning technologies and the widespread deployment of learning management systems (LMSs), WUM methodologies could be used to evaluate learning data. Nonetheless, specific properties of LMSs facilitate the process. For instance, user identification is uncomplicated because all learning platforms demand that a student log in (Marquardt et al., 2004; Munk & Drlík, 2011). Moreover, the flexibility of online learning lets students organize their studies in their own way. Hence, since a student can switch from one section of the course to another and access and return to learning material many times, there is no navigational path that can directly identify a session. Similarly, modern LMSs such as Moodle store student interaction information in their relational databases; as a result, conventional WUM analysis of LMS data does not need an examination of web server logs, thereby simplifying the data cleaning process. As a consequence, although a variety of heuristics can be used to segment each user event log into sessions, in the LA and educational data mining (EDM) fields, sessions are usually derived from log data as continuous sequences of events occurring within a threshold of each other.

Common approaches set the threshold, usually called the session timeout threshold (STT) (Spiliopoulou et al., 2003), to 15 minutes (Munk & Drlík, 2011), 30 minutes (Ba-Omar et al., 2007; Khan & Pardo, 2016; Jovanovic et al., 2017; Matcha et al., 2020), or 60 minutes (Y. Lee et al., 2020; S. Lee et al., 2021). However, to the best of our knowledge, no optimal cut-off strategy has been identified yet (Kovanović et al., 2015). Indeed, despite being straightforward, these time-oriented heuristics are not data driven; they have been shown to be somewhat arbitrary, overestimating or underestimating durations (Nguyen, 2020); they are relatively crude: setting a constant time threshold does not account for the fact that the duration of some activities depends on the activity at hand (for instance, writing a post on a forum or completing an assignment usually takes longer than viewing an entry in the glossary). Therefore, limiting the duration to a fixed threshold necessitates making assumptions that can either decrease or increase the time spent on a task (Rotelli & Monreale, 2022). Moreover, the flexibility of online learning lets students organize their studies in their own way. Hence, since a student can switch from one section of the course to another and access and return to learning material many times, no navigational path directly identifies a session.

Students act differently at different times. Therefore, the initial step in investigating the temporal behaviour of online learners could be identifying the most representative time segmentation of student temporal learning habits. Earlier studies have proposed the following classification of time segmentation strategies: segmentation into pedagogical units (such as courses, class meetings, or lessons) (Molenaar & Friend Wise, 2022; Chiu et al., 2014) and time units (also called clock-based) using a universally agreed upon time division, for instance, a week (Alvarez Sainz et al., 2019; Y. Lee, 2018), a weekday/weekend (Du et al., 2019; Sher et al., 2022; Panzarasa et al., 2015), or an hour of the day (Sher et al., 2022; Panzarasa et al., 2015).

At the same time, there is a second classification for the time segmentation strategies based on how the time-windows are selected. The time-windows may be determined before analysis (without considering the data) or by analyzing the data at hand. In the first case, the time segmentation can be assigned by designating time periods *intuitively* (Sher et al., 2022) or depending on personal experience and perspectives (Du et al., 2019). Although this method allows for speedy answers, it has

the drawback that the outcome is arbitrary and depends on individual assumptions. Considering the time segmentation done with data analysis, we can list, for example, time-windows determined based on the prevalence of low versus high cognitive activities (Chiu et al., 2014).

In our work, starting from the definition of a *task*, which relies on the definition of an *activity*, i.e., a tuple composed of course/area, module, and action (as we will further clarify in Section 5.1), we propose a data-driven methodology for supporting time-window identification. However, it is important to clarify that *task* and *activity* can assume different meanings. For instance, in the field of learning design (Goodyear, 2005; Goodyear & Retalis, 2010), the terms *activity* and *task* have different meanings and are even used in the opposite way: a concrete task is a part of a larger conceptual student activity, i.e., "the mix of psychological and physical activity through which each student responds to their current task." In this paradigm, activity is defined by task (concrete action, like writing an essay or answering a question), people (social context, like individual or collective work), and artifacts/tools (resources and materials, like books and LMSs). Unfortunately, if we only have access to log data, it is impossible to adopt learning design definitions of *task* and *activity* because we cannot capture the "people" aspect, i.e., we cannot determine whether a log was recorded during individual, group, or class study. In addition, the artifacts/tool dimension would become meaningless because its value would be "LMS" for all the activities. Another example of confusion might be related to the Education Modeling Language (EML) and, in particular, its specification called IMS Learning Design (IMS LD) (IMS Global Learning Consortium and others, 2003; Heyer et al., 2007). In this setting, the *activity* refers to what a certain role does and describes what that role is and what environment is available to it within the act. Since in this paradigm the explicit roles specified are those of learner and staff, and we are only interested in the activities performed on the platform by students, all our entries would have the same role: "learner." The environment part, in turn, is defined by two parts: the learning object (a resource used to perform learning activities) and the service (a communication facility). As will be discussed in Section 5.1, in a course, a student can be involved in various tasks, but only a subset of those could be described by the learning object (only the ones that, by our definition, are related to "quality learning"). We didn't want to reduce our work to the study of those particular activities, so it would not be possible to adapt the IMS LD definitions of activity. Concerning the definition of a *task*, to our knowledge, the task is not formally defined in the IMS LD specification.

3. Learning Data

To present our work, we refer to a dataset extracted from a Moodle platform that describes the online learning activities of students of a blended postgraduate master's program that combines online educational materials and face-to-face classes. Moodle courses were all organized and structured to complement face-to-face lectures and laboratories and to provide additional practice activities and materials for student online self-learning beyond face-to-face classes. Thus, our dataset consists exclusively of the sequence of event logs that represent interactions conducted online by students on the Moodle platform. Moodle courses, designed and structured to complement face-to-face lectures, were all organized with various resources[1] (including Book, URL, File); items that can be browsed online or downloaded; and activities[2] (including Lesson, Glossary, Assignment, Quiz, H5P, Feedback, Forum), more advanced modules involving student interaction to support students in learning.

We consolidated our data using the methodology suggested in Rotelli and Monreale (2022) by collecting Moodle log data at the course and at the site level to have a complete overview of user interactions. Since a user who logs in to the online environment can browse the platform and take numerous actions without having to be enrolled in a course (such as viewing the Dashboard, checking the Calendar, sending a Message[3]), we first collected the dataset of all platform logs. Second, we extracted the timestamps directly from the Moodle database and added them as an additional field to the dataset. Third, we extracted the course logs for each of the platform courses and added a new field identifying the course to each dataset; fourth, we merged all course datasets into one. Finally, we joined these data to the site-level table, yielding all logs relevant to user actions while guaranteeing that the time information was preserved. An integrated Moodle log record is supplied with the features described in the following.

Time: the system's sequential time in a standard format;

Course_Area: course or area of the platform where actions were carried out;

Username: the user performing the action;

Recipient: the eventual recipient of the action (for instance, the user who receives a message);

[1] MoodleDocs—Resources—https://docs.moodle.org/401/en/Resources.
[2] MoodleDocs—Activities—https://docs.moodle.org/401/en/Activities.
[3] We refer to the official Moodle documentation for all elements of Moodle's modules: MoodleDocs—https://docs.moodle.org.

Context: the context within the platform[4];

Component: the module type (e.g., *Wiki*, *Page*, *File*, *Url*, *Quiz*[2]);

Event name: the type of action performed on the module (such as *viewed*, *deleted*, *updated*, *created*, *submitted*);

Description: the description of the event;

Origin: the selected *sources*; and

IP address.

An example of an integrated Moodle log record is illustrated in Table 1.

Table 1. Moodle Integrated Log Record

ID	Time	Course_Area	Date&Time	Username	Recipient
402267	1615214006	ICT	2021-03-08T15:33:26	USER 37	-
Context	**Component**	**Event name**	**Description**	**Origin**	**IP address**
Lesson: Introduction	Lesson	Question viewed	The user with id '37' has viewed the Numerical question with id '484' in the lesson activity with course module id '641'	web	109.52.45.53

For data cleaning and preparation, we also refer to the reference paper Rotelli and Monreale (2022). After data consolidation, the number of logs is 313,901.

4. Problem Formulation

In this paper, we tackle the problem of identifying adequate time-windows that could be used to investigate the temporal behaviour of students in OLEs and to adapt analysis techniques to a given dataset. With respect to past research that identifies time periods intuitively or based on personal experience and viewpoints (Section 2), we want to design a process that supports the identification of time periods using a data-driven approach to make the process as objective as possible.

To address the identification of the most representative time segmentation of student temporal learning habits, first, we propose a novel perspective for identifying sessions based on the *duration of inactivity*. Second, we propose an algorithm that assists researchers in calculating the most suitable threshold for the duration of inactivity employed in a session definition. Finally, starting from the identified sessions, we identify time-windows using data-driven methods. To this end, as in the case of dashboards that enable insightful real-time analysis (Macfadyen & Dawson, 2010), we designed and developed a visual tool that assists data scientists and researchers in determining the optimal settings for session identification and locating suitable time-windows.

5. Session Identification

This section formally introduces the proposed methodology for session identification. In particular, we first define the notion of *session* in an online learning platform and then we provide a data-driven approach for extracting these sessions.

5.1 Defining Study, Course, and Learning Sessions

The input of our analytical process is an *activity dataset* $\mathscr{D}$, i.e., a collection of *student activity sequences* $\mathscr{D} = \{A_1, \ldots, A_m\}$, where each A_x is a sequence of events that describe the activities of a student x in an OLE.

Definition 1 *(Student Activity Sequence) The activity sequence of a student x is a chronologically ordered sequence of pairs* $A_x = \langle (a_1,t_1), \ldots, (a_n,t_n) \rangle$*, where* $a_i \; \forall i \in [1,\ldots,n]$ *describes the activity that the student performs in an OLE and* t_i *denotes the activity timestamp such that* $\forall\ 1 \leq i < n,\ t_i \leq t_{i+1}$.

We aim to identify the time-windows to analyze student temporal behaviour; thus, we first have to measure the time students spend online. To achieve this objective, we have to determine each activity's duration.

[4]MoodleDocs—Context—https://docs.moodle.org/39/en/Context.

Definition 2 *(Activity Duration) Given two consecutive activities (a_i,t_i) and (a_{i+1},t_{i+1}) of a student activity sequence A_x (with $t_i \leq t_{i+1}$), the duration d_i of the activity a_i is the difference between its timestamp and that of the next activity in the sequence, i.e., $d_i = t_{i+1} - t_i$.*

For each activity in A_x we can replace the timestamp t_i with the duration d_i to derive the *student duration activity sequence*.

Definition 3 *(Student Duration Activity Sequence) Let $A_x = \langle (a_1,t_1),\ldots,(a_n,t_n)\rangle$ be the activity sequence of a student x. The duration activity sequence A_x^d is derived by replacing t_i with the information about the duration d_i of the activity a_i, i.e., $A_x^d = \langle (a_1,d_1),\ldots,(a_n,d_n)\rangle$.*

In traditional presence-based courses, Carroll (1963) distinguishes *time-on-task* and *time-off-task* as the time spent learning and the time spent in activities other than learning, together forming the *elapsed time*. *Quality learning* has been defined as "learning in which learners are provided with the ability to effectively learn and retain skills and knowledge gained" (Thindwa, 2016). The distinction between *on-task* and *off-task* activities is evident in face-to-face courses.

In online environments, students can participate in a course or browse the platform without engaging in course activities. Hence we can differentiate between the *course activity* taken on courses and the *platform activity* taken on platform areas, together forming the *on-task activity*. However, the analysis of the activity duration distributions, most of which are characterized by very long tails, highlights that in different cases the duration of some actions vastly deviates from other observations in the data. This confirms the analytical results of previous research studies (Kovanović et al., 2016; Nguyen, 2020; Rotelli & Monreale, 2022) on the very long duration of some learning activities. We cannot assume that the total time students are online is entirely dedicated to interaction. While connected, they can also be inactive, take some breaks, or browse other websites, i.e., they can be *off-task*. As a result, looking closely at our definition of *activity duration* (Definition 2), we observe that it represents Carroll's *elapsed time*. This duration includes the time spent on *on-task* activities and the time spent on *off-task* activities.

Each course is provided with various learning *modules* (e.g., files, videos, quizzes) on which students can perform a range of *actions* (e.g., view, create, update). Similarly, students can interact with the platform's *modules* (e.g., dashboard, calendar, messaging). As a result, each *activity* a_i can be described by the course or platform area γ_i, the module μ_i, and the type of action α_i, i.e., $a_i = (\gamma_i,\mu_i,\alpha_i)$. Hence, the duration activity sequence of a student x (Definition 3) can be represented by $A_x^d = \langle (\gamma_1,\mu_1,\alpha_1,d_1),\ldots,(\gamma_n,\mu_n,\alpha_n,d_n)\rangle$. For instance, given a student, their activity sequence can be $\langle (Course\ A, File, viewed, 340), (Course\ A, Lesson, started, 5), (Course\ A, Lesson, question\ viewed, 56),\ldots,(Course\ A, Book, chapter\ viewed, 87)\rangle$.

To deal with off-task activities, we started from the assumption that the time students spend on task may vary from activity to activity and may differ from course to course. For instance, watching a video can take longer than reading a text; answering multiple-choice questions in the quiz of course A can take less time than answering short-answer questions or essay questions in the quiz of course B. Therefore, we adopt the approach proposed by Rotelli and Monreale (2022), who estimate the duration value for each course/area γ_i, for each type of action α_i taken on each specific module μ_i, separately. However, the time student A takes to complete the activity might differ from the time student B takes to finish the same activity. To also account for each student's individual learning time requirements, we propose to adapt their methodology to individual student sequences and estimate the duration value for each student x. Consequently, to calculate the estimated activity duration, we need to extract the subsequence related to the same course, the same module, and the same action, from each student duration activity sequence A_x^d. More formally, we have the following definition.

Definition 4 *(Subsequence) Given the student duration activity sequence $A_x^d = \langle (\gamma_1,\mu_1,\alpha_1,d_1),\ldots,(\gamma_n,\mu_n,\alpha_n,d_n,)\rangle$ of a student x, $S_x = \langle (\gamma'_1,\mu'_1,\alpha'_1,d'_1),\ldots,(\gamma'_h,\mu'_h,\alpha'_h,d'_h)\rangle$ is a subsequence of A_x^d or is contained in A_x^d ($S_x \preceq A_x^d$) if there exist integers $1 \leq i_1 < \cdots < i_h \leq n$ such that $\forall j \in [1,\ldots,h]$, $(\gamma'_j,\mu'_j,\alpha'_j,d'_j) = (\gamma'_{i_j},\mu'_{i_j},\alpha'_{i_j},d'_{i_j})$.*

Since most of the duration distributions tend to be right skewed, as also illustrated in previous work (Kovanović et al., 2016; Nguyen, 2020; Rotelli & Monreale, 2022), for each type of action α taken on each specific module μ of each course/area γ of each student x, we estimate the duration by calculating the median value and the median absolute deviation (MAD) of the corresponding values $d_1,\ldots,d_p$. Therefore, the estimated duration δ_i is equal to the median if d_i is bigger than the threshold, derived as the sum of the median and the MAD. In every other case δ_i is equal to d_i, i.e., $\forall i \in [1,\ldots,p]$, δ_i = median if $d_i >$ median+MAD; $\delta_i = d_i$, otherwise.

The difference between the activity duration d_i and the estimated activity duration δ_i represents the time students are inactive between two consecutive activities, i.e., the *estimated inactivity duration* $\overline{\delta_i}$, i.e., $\overline{\delta_i} = d_i - \delta_i$. We can now enrich the student duration activity sequence A_x^d with the estimated duration of each inactivity, i.e., we can derive the *student temporal sequence*.

Definition 5 *(Student Temporal Sequence) Let* $A_x^d = \langle(\gamma_1,\mu_1,\alpha_1,d_1),\ldots,(\gamma_n,\mu_n,\alpha_n,d_n)\rangle$ *be a student duration activity sequence. The student temporal sequence* $\mathscr{T}_x = \langle(\gamma_1,\mu_1,\alpha_1,d_1,\overline{\delta_1}),\ldots,(\gamma_n,\mu_n,\alpha_n,d_n,\overline{\delta_n})\rangle$ *is the student duration activity sequence* A_x^d *enriched with the estimated inactivity duration* $\overline{\delta_i}$.

For instance, the student temporal sequence of the previous example will be $\langle(Course\ A, File, viewed, 340, 230),(Course\ A, Lesson, started, 5, 0),(Course\ A, Lesson, question\ viewed, 56, 31),\ldots,(Course\ A, Book, chapter\ viewed, 87, 29)\rangle$.

We are interested in estimating the time a student spends *on task*, as Carroll (1963) defined. However, the "time-on-task" can assume different facets in the LA research field. Several studies pinpointed the *time-on-task* as the duration between two consecutive clicks, but it has also been defined as the number of clicks in a time-window period, the time spent on course activities, and the time spent on a learning task (Rotelli & Monreale, 2022). By analyzing the instances of the student temporal sequence $\mathscr{T}_x$, we can establish that each record in the sequence does not reflect a task but rather an activity a_i which can be described by an action α_i performed on a module μ_i in a course γ_i, i.e., $a_i = (\gamma_i,\mu_i,\alpha_i)$. As a consequence, we adopt the definition of Romero and colleagues (2014), who defined a *task* as a sequence of interactions of a user on a given module, where each interaction is a single record in the log file. We aggregate all contiguous actions α related to the same module μ in a course or area γ of the platform.

Definition 6 *(Contiguous Subsequence) Given the temporal sequence* $\mathscr{T}_x = \langle(\gamma_1,\mu_1,\alpha_1,d_1,\overline{\delta_i}),\ldots,(\gamma_n,\mu_n,\alpha_n,d_n,\overline{\delta_n})\rangle$ *of a student x, a contiguous subsequence* G_x *of* $\mathscr{T}_x$ ($G_x \preceq^c \mathscr{T}_x$) *with length* $k \leq n$ *is a sequence of k contiguous activities of* $\mathscr{T}_x$, *i.e.,* $G_x = \langle(\gamma_j,\mu_j,\alpha_j,d_j,\overline{\delta_j}),\ldots,(\gamma_{j+k-1},\mu_{j+k-1},\alpha_{j+k-1},d_{j+k-1},\overline{\delta_{j+k-1}})\rangle$, *for* $1 \leq j \leq n-k+1$.

Now we can formally define a task as follows.

Definition 7 *(Task) Given the temporal sequence* $\mathscr{T}_x = \langle(\gamma_1,\mu_1,\alpha_1,d_1,\overline{\delta_i}),\ldots,(\gamma_n,\mu_n,\alpha_n,d_n,\overline{\delta_n})\rangle$ *of a student x, a task T is a contiguous subsequence of* $\mathscr{T}_x$ *concerning the same course or platform area and the same module, i.e.,* $T = \langle(\gamma_j,\mu_j,\alpha_j,d_j,\overline{\delta_j}),(\gamma_{j+1},\mu_{j+1},\ \alpha_{j+1},d_{j+1},\overline{\delta_{j+1}}),\ldots,(\gamma_l,\mu_l,\alpha_l,d_l,\overline{\delta_l})\rangle$, *such that* $T \preceq^c \mathscr{T}_x$, $\gamma_j = \gamma_{j+1} = \cdots = \gamma_l$, *and* $\mu_j = \mu_{j+1} = \cdots = \mu_l$.

Referring to the previous example, a task could be $T = \langle(Course\ A, Lesson, started, 5, 0),(Course\ A, Lesson, question\ viewed, 56, 31)\rangle$. Given the duration of each action in a task T, the total time spent on T, namely the *elapsed time*, is derived from the sum of all durations, i.e., $D = \sum_{i=j,j+1,\ldots,l} d_i$. Accordingly, to calculate the *time-off-task*, we have to sum up the inactivity duration value of every task action, i.e., $\overline{\Delta} = \sum_{i=j,j+1,\ldots,l} \overline{\delta_i}$. The task T performed on the course or platform area γ on the module μ can now be represented more compactly, i.e., $T = (\gamma,\mu,D,\overline{\Delta})$, i.e., $T = (Course\ A, Lesson, 61, 31)$.

We define a *study session* as the collection of all tasks performed online between the login and the last interaction before inactivity. We assume that if the *time-off-task* is greater than zero, the student was distracted; hence, the study session ends. Nevertheless, given that the *time-off-task* $\overline{\Delta}$ is the sum of the estimated inactivity durations and that a brief pause does not necessarily indicate that a student was wholly distracted, we use the STT θ. Task T is considered to be the last task of the session if the *time-off-task* is higher than the STT.

We can now formally define a *study session* as follows.

Definition 8 *(Study Session) A study session* $\mathscr{S}$ *is a collection of tasks* $\mathscr{S} = \{T_1,\ldots,T_z\}$, *where* $\forall i \in [1,\ldots,z-1]$ *the time-off-task is less than or equal to the STT* θ, *i.e.,* $\overline{\Delta_i} \leq \theta$, *while the time-off-task of* T_z *is bigger than that threshold, i.e.,* $\overline{\Delta_z} > \theta$.

Consequently, the student temporal sequence $\mathscr{T}_x$ (Definition 5) can be described as a collection of study sessions, i.e., $\mathscr{T}_x = \{\mathscr{S}_1,\ldots,\mathscr{S}_{n'}\}$.

As previously mentioned, when students are connected, they can interact in a course or browse the platform to look for information, view the dashboard, send messages to other students, or do something unrelated to any course. As a consequence, "being online" does not imply interacting in a course. Moreover, students can enroll in several courses. Hence, to examine student activity in different courses, we have to extract the contiguous subsequence of events related to a specific course γ. Therefore, the *course session* is the collection of all contiguous tasks performed in the same course.

Definition 9 *(Course Session) A course session* $\mathscr{C} = \{T_1,\ldots,T_w\}$ *is a contiguous subsequence of a study session* $\mathscr{S}$ *related to the same course* γ ($\mathscr{C} \preceq \mathscr{S}$), *where the session is ended if the student moves to another course or becomes inactive (i.e.,* $\overline{\Delta_w} > \theta$).

Likewise, in a course, a student can be involved in various tasks, but only a subset of those are related to "quality learning," i.e., activities with intentional and purposeful student involvement that enable effective learning and retention of acquired skills and knowledge (Thindwa, 2016). Not every activity has a specific learning involvement. As a consequence, we consider quality learning activities to be those that necessitate an active role in the learning process, i.e., facilitate the acquisition of skills and knowledge. For instance, the *Attendance* module, unlike *Lesson*, does not involve any "active learning," since the student action is limited to taking attendance and has no effect on their academic performance or knowledge acquisition. On the other hand, there are other activities, like *Quiz*, that require students' active participation and influence academic performance, during which a student can "effectively learn, retain skill and knowledge" (Thindwa, 2016). Specifically, in our dataset, we consider the following modules being related to quality learning:

File: slides, PDF files, course materials, etc.;

Glossary: collection of definitions related to the course materials;

URL: link to a website, lesson recordings, or online file;

Assignment: module that allows students to submit their work for grading;

H5P: interactive activities designed to help students better comprehend the theoretical concepts;

Quiz: from simple tests with multiple-choice questions to complex tasks with detailed feedback; and

Lesson: "branching" exercises in which students can access content and questions and then be directed to specific pages based on their responses[5].

Other alternatives are possible based on individual needs. Thus, we can extract the *learning session* as the contiguous subsequence of the course session with "quality learning" tasks.

Definition 10 *(Learning Session) A learning session $\mathscr{L} = \{T_1, \ldots, T_v\}$ is a contiguous subsequence of $\mathscr{C}$ related to those tasks associated with quality learning ($\mathscr{L} \preceq \mathscr{C}$), where the session is ended if the student moves to a task that differs from learning or becomes inactive ($\overline{\Delta}_v > \theta$).*

The above three definitions of *session* depend on the STT value θ. In the following, we propose a data-driven approach to derive the value that best fits the data under analysis for each type of session.

Before discussing the main ideas of our approach, it is crucial to stress some key points. First, differing course structures can result in the collection of different sequences of activities (Definition 1). Second, although we present our research using a Moodle dataset, our methodology generally applies to any OLE. A tuple composed of a learning activity (course, component, and event name, as illustrated in Table 1) and a timestamp is generically derivable from Moodle and any OLE that can trace back user log data. For instance, Coursera log data record the Learning Material (μ) (corresponding to Moodle's *component*) in the course (γ), the Timestamp (t), and the EventSubType (α) (corresponding to Moodle's *event name*). Once the duration (d) is computed, the tasks (T) (Definition 7) can be identified; thus, the formal definitions we propose can be transferred and applied to this specific OLE and any other that can trace back user logs to the tuple (course, module, action, duration). Third, in Moodle, all actions related to the online environment browsing (e.g., viewing the *Dashboard*), as well as the *user-user* relationship (for instance, sending and reading private messages via the platform Messaging system rather than the public *Chat* module that can be added to a course), or involving student self-awareness (i.e., aspects related to the user profile) are logged not at the *course* level, but at the *platform* level. Since our definition relies on the fact that quality learning occurs within courses, and despite the fact that a "messaging" action can be an action directly related to "learning," Moodle's logging system does not store messages in course logs; consequently they cannot be considered learning modules.

5.2 Estimation of the STT for Session Identification

A session ends if $\overline{\Delta} > \theta$, i.e., if the time-off-task is bigger than the STT, regardless of the session type. The possibility of setting a custom STT is crucial to visualize and understand its impact on the outcomes. However, selecting the ideal parameters for various objectives is not a trivial task, and the STT has proved to be the most challenging (Section 2). To assist in the selection of an STT, we propose an algorithm that recommends the most suitable STT based on the input parameters and data. In the following, we discuss the approach used to determine the STT for study, course, and learning session.

[5] MoodleDocs—Managing a Moodle course—https://docs.moodle.org/39/en/Managing_a_Moodle_course.

STT for study sessions. To determine the optimal STT for study sessions, we observe whether students return to the same area of the platform after a period of inactivity or begin working on a different one. We assume that if students interact in a different area after a period of inactivity, this could be interpreted as a signal for a new session. In particular, we analyze student behaviour to derive the STT using the following logic: we assume that the time-off-task cannot exceed the maximum STT set. For each minute of the inactive period between zero and max STT, we record the number of times students remain in the same area or switch to a different one. Then, we analyze one minute at a time, looking at the likelihood that the inactivity period will be followed by one of the two different types of behaviours. To calculate this probability, we proceed as follows: while we cannot compare the sizes of the two behaviours for every minute since they are not of equal size, we use their densities over the whole number of occurrences. A frequency density measures the frequency per unit for each label datum. To compute it, we divide the frequency by the label size, i.e., the occurrences of a single event by the number of possible events. In our data, for minute 0 to minute 1, students remain on the same course 14,113 times (frequency) out of 28,939 events (label size) of this behaviour. The density for minute 0 to minute 1 is 0.49, which is obtained by dividing 14,113 by 28,939. Students switch labels in the first minute 6,782 times out of 20,409 occurrences. The density is 0.33 (equal to 6,782/20,409). Therefore, the probability of continuing in the same area is $0.49/(0.33+0.49) = 0.6$, while the probability of switching is 0.4. Because in this case the probability that students remain in the same area is greater than the probability of students switching, we could establish that for minute 0 to minute 1, students tend to stay in the same area after inactivity.

We may simply take the greater of the two values to determine which behaviour exhibits the greater trend. Instead, to get a clearer view of student behaviour, we subtract the probability of students switching from the probability of students staying. This value is referred to as *behaviour deviation*. If it is greater than zero, it suggests that students tend to stay in the same area with this duration of inactivity. Otherwise, if it is less than zero, students tend to switch areas of the platform.

In Figure 1, where the x-axis represents the inactivity period (expressed in minutes) and the y-axis represents the behaviour deviation (i.e., probability of staying minus probability of switching), we observe that the data follow the function $y = a * ln(x) + b$, with a constantly decreasing likelihood of staying in the same area. In this and the next two figures, we have limited the maximum STT to 60 minutes, since this is the maximum STT determined by previous research (Section 2).

As previously indicated, we assume that if students switch areas after a period of inactivity, this widespread behaviour could be interpreted as the end of the session and used as a threshold. Consequently, we are interested in identifying the point at which this behavioural switch happens. To address this problem, we propose performing a curve fitting, i.e., constructing a curve of the form $y = a * ln(x) + b$, to find the best fit to a given set of behaviour deviation points. We consider the point at which the curve intersects the x-axis as the STT, which we recommend. For example, in Figure 1, the red line, which represents the best fit for our deviation points (the function is located top right), intersects the x-axis at 3.9 minutes, which we consider the recommended STT.

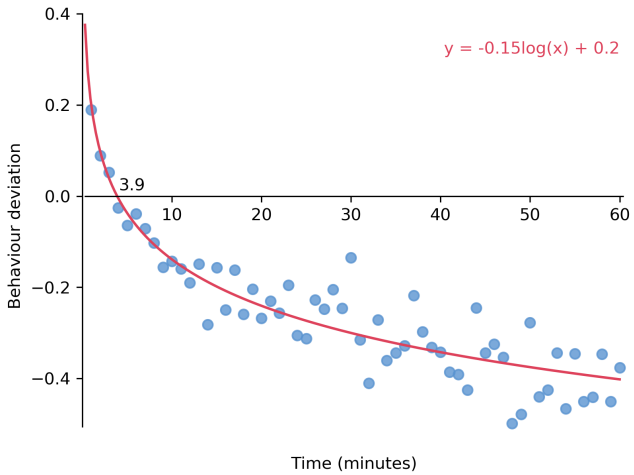


Figure 1. Example of behaviour deviation distribution to identify the STT for study sessions.

STT for course sessions. To determine the recommended STT for course sessions, we observe whether students switch courses or enter a period of inactivity and whether, in the second instance, they return to the same course after the inactivity period. We assume that if students begin working on another course, this could be interpreted as a signal for a new session. To derive the STT for course sessions, we apply a similar method as for study sessions: the interval between zero and max STT is analyzed minute by minute, and the received values are curve-fitted to receive the precise point of behavioural shift.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

Figure 2 illustrates the trend of behaviours for course sessions. We aggregated these values on the assumption that switching courses is indicative of a new session. Notably, the course changeover can also be preceded by a period of inactivity, meaning that students might remain inactive prior to switching courses. This period of inactivity serves as the value for the course switch. Typically, students continue to work in the same course for up to 3.2 minutes of inactivity. During inactive times longer than 3.2 minutes, students typically switch courses.

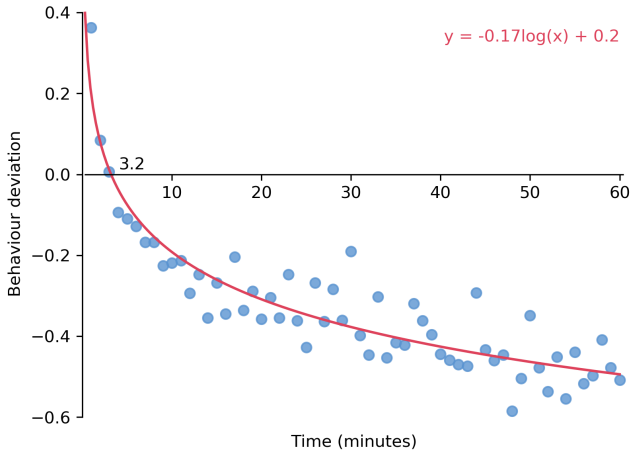


Figure 2. Example of behaviour deviation distribution to identify the STT for course sessions.

STT for learning sessions. To determine the recommended STT for learning sessions, we examine whether students who have interacted in a quality learning module (Section 5.1) keep working on this type of module after a period of inactivity or begin working on modules unrelated to quality learning. To derive the STT for learning sessions, we work similarly to the course sessions' STT identification. Figure 3 illustrates the trend of behaviours for learning sessions. Up to 3.7 minutes of inactivity, students tend to continue to work on the quality learning modules, whereas after this period, they switch to non-quality learning modules.

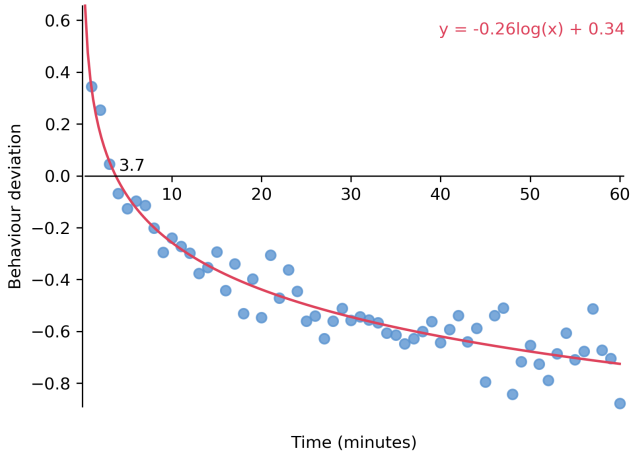


Figure 3. Example of behaviour deviation distribution to identify the STT for learning sessions.

It is worth noting that this trend has been observed in all experiments we conducted since the descending curve describes the behaviour of the students, i.e., when inactivity increases, students typically switch activities or courses. To confirm whether this trend can be extrapolated to other datasets, additional research is required.

The STT θ, whether recommended or manually set, serves as a delimiter for all tasks T of the session under consideration. We have already mentioned that the time students spend on-task may vary from activity to activity and course to course. Consequently, we also offer the option to set a separate threshold for specific modules. To recommend a threshold for specific modules (one or more), only the occurrences in which the specified module occurred before the inaction are taken into account

for the analysis. As a result, selecting one or few modules could result in insufficient data and statistically irrelevant results. This is an important aspect that an investigator has to take into account when setting the tool parameters.

The estimation of the STT is a fundamental step for the time-window identification described in the next section. However, it could also be very useful for other data-driven investigations. For example, it can help explore how different modules (μ) keep the students busy longer in a session. By estimating the STT for study sessions for each module of our dataset, we identified the occurrences of ending modules, which are reported in Table 2. Apart from the Logout module, a module is considered an *ending module* if its time-off-task exceeds the STT. Table 2 displays, for each module of the study session, the number of occurrences of that module, the number of times it appears as an ending module, and the percentage of occurrences as an ending module with respect to its total occurrences in the entire dataset.

Table 2. Module Analysis of the Study Session to Determine Which Modules Are Typically Followed by a Period of Inactivity

Module	# Occurrences	# Session-end	% Session-end
Logout	845	845	100.00%
Messaging	954	476	49.90%
File	8467	3437	40.59%
Assignment	2598	907	34.91%
URL	3284	1098	33,43%
Participant_profile	1496	459	30.68%
Dashboard	23652	7123	30.12%
Folder	2398	721	30.07%
User_profile	801	225	28.09%
Page	924	245	26.52%
Grades	1733	450	25.97%
Choice	396	92	23.23%
Course_home	44128	7829	17.74%
H5P	1362	241	17.69%
Notification	169	28	16.57%
Attendance	6012	947	15.75%
Quiz	11942	1851	15.50%
Glossary	765	100	13.07%
Forum	2715	345	12.71%
Lesson	9875	1180	11.95%
Book	496	52	10.48%
Site home	517	41	7.93%
Feedback	301	21	6.98%
Courses list	1247	84	6.74%
Database	144	9	6.25%
Label	138	8	5.80%
Reservation	327	18	5.50%
Badge	112	5	4.46%
Survey	318	14	4.40%
Login	7764	158	2.04%
Tag	10	0	0.00%

The results reveal that students are particularly inactive after interacting with *Files*[6], *Assignments*[7], and *URLs*[8]. When analyzing the module *File* in greater depth, we must consider the possibility that some students could download a file and read it offline, while others could read it online. In both cases, since the platform does not record any additional activity (Definition 1, i.e., no log is recorded), the inactivity period does not necessarily indicate that the student is distracted but rather that the student is performing no other actions on the platform. Regarding *URLs*, the case is similar to *Files*. In Moodle, a URL is a hyperlink that leads to an external website or an external or internal file. If we take as an example a link to a video, when a student watches a video, no more action is performed. Therefore, the duration of the action of watching will undoubtedly exceed the STT detected between module changes. This does not imply that students are distracted; they are simply not interacting with

[6]https://docs.moodle.org/401/en/File_resource.
[7]https://docs.moodle.org/401/en/Assignment_activity.
[8]https://docs.moodle.org/401/en/URL_resource.

the platform. Similarly to *File*, *Folder*[9] involves either downloading or displaying files for reading on the platform. Also in this case, the inactivity period does not necessarily indicate that the student is distracted. Concerning the assignment, to understand if the student is distracted or not, it would be necessary to perform a detailed analysis of the actions, i.e., to determine if there is an action that could be identified as conclusive of the task before inactivity (in particular *Assignment submitted*). This may suggest that the student took a break after completing the assignment. In contrast, if the task is not completed (for instance, *Assignment viewed*) and is interrupted by inactivity, this may indicate that the student became distracted.

After these first modules, there is a series of platform modules in the table that do not involve *quality learning* but rather facilitate navigation within the learning environment, such as *User_profile* or *Course_home*. In these cases, the period of inactivity could be attributed to a student leaving the platform without logging out after browsing.

Beyond the Assignment module, the most frequent quality learning modules are *H5Ps*[10] (i.e., interactive activities), *Quizzes*[11], and *Glossaries*[12]. However, we highlight that a low percentage value (17.74%, 15.50%, and 13.07%) characterizes them as ending modules. These low values lead us to infer that students are focused when interacting with this type of module.

6. Data-Driven Time-Window Identification

The foundation of our methodology is the idea that the data of interest vary based on the objectives of the investigation. The investigation might only be interested in specific courses or particular time periods, in all the platform activities or only in the activities concerning learning, etc. These choices affect the resulting sessions, which in turn influence the time-windows we identify depending on the sessions. These considerations led us to conclude that developing a visual tool would be the simplest way to approach this issue. Providing an interactive way of selecting the data and parameters to consider would be a straightforward way to build the plot through which time-windows could be identified (Section 6.2).

To create the visual tool, we used Dash, a Python framework[13], which provides a declarative Python interface for developing full-stack, interactive web applications and has been used to visualize data in several domains (Hossain et al., 2019; Britzolakis, 2020). The Dash applications are made up of two blocks: layout and callbacks. The layout is responsible for the visual part of the tool, and to give the desired look, it is possible to use Dash HTML[14] and Dash Core components[15] and style them with CSS sheets. The callbacks are used to program the application's interactive features, whose whole source code is available on GitHub[16].

We made the tool accessible online[17] as well a sample of the dataset logs to provide the opportunity to test and to interactively customize the settings for specific requirements (Section 6.1), such as setting an STT, and visualize in real time how each customization affects the time-window identification plot. To fulfill this step, we needed to deploy the application to a server. We used the *Render* cloud platform[18], which natively supports Python web applications and offers free deployment.

Figure 4 depicts the tool's appearance. We can observe a distinct divide between the two parts: in the top panel, data and parameters can be selected, whereas the bottom panel shows the plot that is used to define the time-windows, as we illustrate in the following sections.

6.1 Parameter Settings for Time-Window Identification

Several factors influence the identification of sessions, and there is no optimal set of parameters that could be considered the gold standard in every situation. Depending on the focus of interest, the kinds of sessions we wish to represent may vary. Thus, to conveniently deal with constantly changing settings, our tool exploits the definitions described in Section 5.1 to enable the selection and customization of parameters whose selection is reflected in the time-window identification plot.

In the following, we list and describe all the parameters that can be set in our tool, as depicted in Figure 4:

- **Log selection.** One or more courses may be chosen. If no courses are selected, the entire log dataset, which consists of course and platform logs, is used to identify sessions.
- **Max STT.** This parameter enables the setting of a maximum STT value, or the retention of the default value of 60 minutes, to be used to identify the recommended STT.

[9] https://docs.moodle.org/401/en/Folder_resource.
[10] https://docs.moodle.org/401/en/H5P.
[11] https://docs.moodle.org/401/en/Quiz_activity.
[12] https://docs.moodle.org/401/en/Glossary_activity.
[13] https://plotly.com/dash/.
[14] DashDocs—Dash HTML components—https://dash.plotly.com/dash-html-components.
[15] DashDocs—Dash Core components—https://dash.plotly.com/dash-core-components.
[16] https://github.com/aleksandraMaslennikova/time_windows_identification.
[17] https://visual-session-identification.onrender.com/. Please note that the tool requires some time to load completely.
[18] https://render.com/.

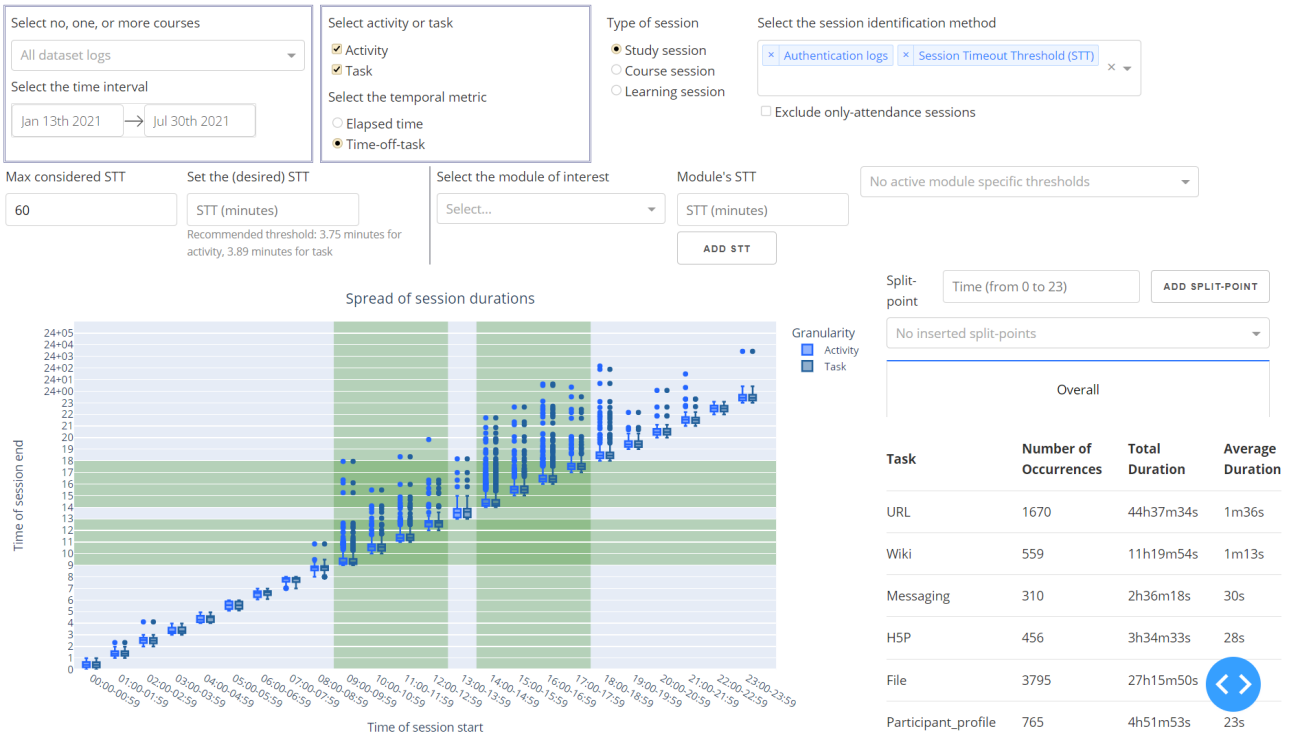


Figure 4. Visual tool's appearance. The top panel allows the user to pick data and parameters, while the bottom panel displays the plot used to define the time-windows on the left and the table on the right displays information on the occurrences of ending modules overall and for each manually inserted time-window.

- **Time interval.** This parameter enables the user to set specific time boundaries or keep those recommended by the tool, based on the selected logs, unchanged.
- **Activity and tasks.** This parameter specifies the record sequence that will be used to identify sessions. In the case of "Activity," the Student Temporal Sequence (Definition 5) is used. It defines student interactions as a sequence of tuples whose elements are the activities defined by a course, a module, and an action. The Task (Definition 7) groups events that belong to the same course and module. Thus, the critical difference for this setting is the granularity. The "Task" gives a broader view, while the "Activity" provides more details, reporting the types of actions performed on the module. The tool enables the selection of one or both granularities.
- **Temporal metric.** This parameter enables the use of the activity duration (Definition 2), i.e., Carroll's *elapsed time*, or the estimated inactivity duration, i.e., the time-off-task $\overline{\delta}$, to identify sessions. Previous research defined the STT for the elapsed time (Jovanovic et al., 2017; S. Lee et al., 2021; Y. Lee et al., 2020; Matcha et al., 2020; Munk & Drlík, 2011; Spiliopoulou et al., 2003). Thus, to provide the opportunity to reproduce these studies, we implemented this feature.
- **Session types.** This category entails identifying the types of sessions to examine: study sessions (Definition 8), course sessions (Definition 9), or learning sessions (Definition 10). The study sessions are interrupted in the event of inactivity; the course sessions are interrupted in case of inactivity or when students switch to a different course; the learning sessions are interrupted if students are inactive or switch to a module unrelated to quality learning. It is essential to emphasize that module usage is course specific and that the definition of quality learning modules differs from course to course.
- **Session identification method.** Two types of identification are available for selection: *authentication logs* and *STT*. The former is based on the platform logins and logouts. The new session begins when the previous log is a logout or when a new login record is encountered. Even though this method is based on the student's concrete behaviours, it does not take into account that in an LMS like Moodle, if users take no action (no requests to the server) during a specific time period (eight hours by default[19]), their session is ended and they are logged out. Within this time-window, users can access the platform many times, even closing the browser. In other words, re-authentication is unnecessary for users accessing the

[19] MoodleDocs—Session handling—https://docs.moodle.org/39/en/Session_handling.

platform within 27,799 seconds (7 h 59 min 59 s). Regarding the second type, we noticed that many users fail to log out before closing the browser. Therefore, we cannot account for cases in which students do not log out because we may run the risk of having unnecessarily long durations. As a result, we implemented the STT.

- **Exclude only-attendance sessions.** This option allows data manipulation based on domain-specific knowledge. Due to the mandatory nature of presence-based lessons, students were required to log in to the Moodle course and take attendance at the beginning of the face-to-face lecture. In some cases, students left the platform immediately after taking attendance to attend the lecture (frequently without logging out by closing the browser). In other cases, they did some activities or exercises on the Moodle course as instructed by the teachers. As a result, we decided to offer the option to remove sessions that comprise solely this mandatory task because they would not contribute any meaningful information and could potentially bias the results.
- **Desired STT.** A specific STT can be set. However, the tool offers an STT recommendation based on the algorithm discussed in Section 5.2, which considers all settings discussed in this section.
- **Module of interest.** This parameter allows the module to which a given STT can be assigned to be selected. We have previously noted that the time students spend on task might vary from activity to activity and course to course. Therefore, the activity type also influences the time-off-task.
- **Module's STT.** To further customize the tool, we also offer the option to specify a different threshold for specific modules. In this case, this module-specific STT will be used instead of the generic one for all sessions when the specified module was the last to occur prior to inactivity. Each time a module-specific STT is set, it is added to the bar on the screen's right. They can be either updated or removed.

6.2 Time-Window Identification

For the time-window identification analysis, after selecting and customizing parameters, we propose to use a plot with 24 boxplots of session duration spread (one for each hour of the day), as illustrated in Figure 5. We recall that a boxplot is a standard technique for illustrating the locality, spread, and skewness of data through their quartiles. The box displays the dataset's minimum, maximum, sample median, and first (25%) and third (75%) quartiles. The whiskers, i.e., the lines extending from the box, represent the variation outside the upper and lower quartiles. Beyond the whiskers, plotted as separate points, outliers depict data that significantly deviate from the rest of the dataset.

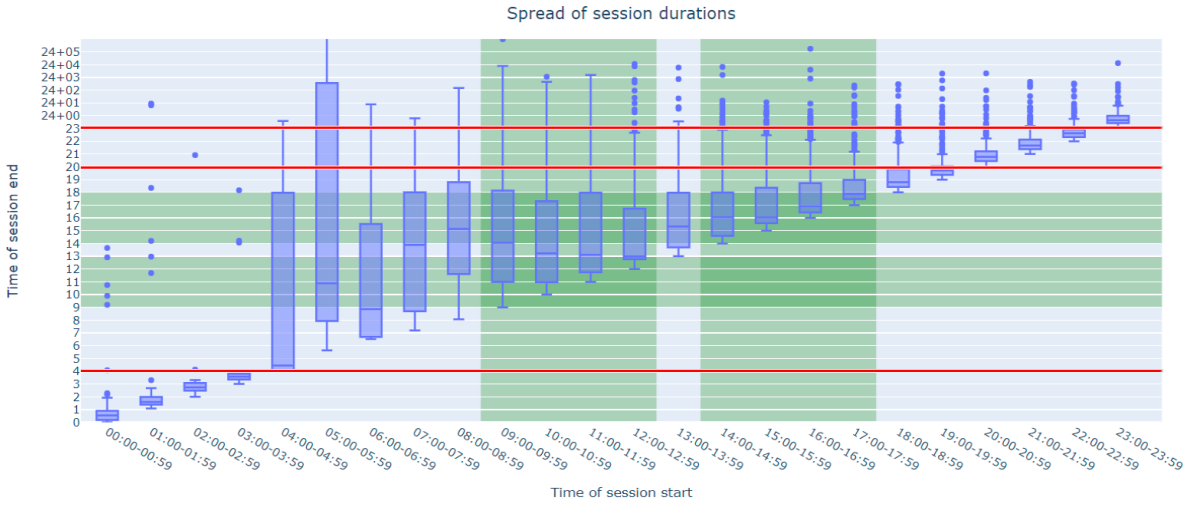


Figure 5. Time-window identification plot. It depicts the plot with 24 boxplots of session duration spread (one for each hour of the day), obtained by selecting and customizing parameters. The *x*-axis illustrates the session's start time, rounded to the hour. The *y*-axis represents the session's end time. The green stripes indicate when students attend lectures and laboratories (domain knowledge—Section 3). The red horizontal lines represent the split-points.

The *x*-axis depicts the session's start time rounded to the hour. The *y*-axis illustrates the session's end time. The green stripes represent the periods when students attend lectures and laboratories (domain knowledge—Section 3). We added them because in our particular case we would like to keep all the Moodle activities that were performed during face-to-face lessons

inside the same time-windows. Clearly, if the dataset does not contain online activities during the lectures, the green stripes can be omitted, since the rest of the proposed algorithm does not rely on them. Hovering the mouse over one of the boxes displays the number of sessions as well as statistical information regarding session duration and ending times (minimum, maximum, the first and third quartiles, the median, the most common activity or task, the most common session-ending task). Thus, whenever the parameter setting is modified, the plot is updated.

The core premise of our approach is to reduce the likelihood of a session being split into two different time-windows. Most sessions should be completely contained within a single time-window rather than divided into two. Optimizing this requirement helps avoid negative analytical issues since sessions split into different time-windows would be analyzed separately. Consequently, we risk missing key behavioural regularities because interconnected events would be analyzed separately. In terms of the plot (Figure 5), this means that the red horizontal line, which we call the *split-point*, used to represent the moment that separates two consecutive time-windows, should ideally not intersect any of the boxplot elements, i.e., box, whiskers, and outliers. Since this ideal separation is hard to obtain in some cases, we attempt to reduce the number of sessions divided into two independent time-windows by minimizing the part of the session assigned to the next time-window. This means we prefer split-points to intersect outliers rather than whiskers or boxes. As a result, for the split, we consider the following preference order: outliers, whiskers, and box. In addition, since the identification of time-windows could also rely on the data domain knowledge, if any (such as, in our case, the timetable of face-to-face lessons, represented on the plot by green stripes), the visual tool could support the investigator in making a more informed final decision.

To demonstrate the utility of our tool, we present a case study that illustrates the power of our time-window identification technique. In particular, in this case study, the goal is to understand *when* and in particular *which moments of the day* students prefer to study by interacting with the learning material available on the online platform. A comparison of our data-driven method with existing work will aid in determining whether the newly proposed definitions of sessions throughout the inactive time are more adaptable to a given dataset than the session definition from prior work.

Once the time-windows have been identified, as displayed in the table at the bottom right part of Figure 4, the tool allows manual entry of the split-point values (such as 12:00, 14:00, 20:00, 23:00) to obtain LA information concerning those particular time-windows, such as the number of occurrences, the total time, and the average time (total time divided by the number of occurrences) for each task.

Case study. In their work, Sher and colleagues (2022) analyzed students' platform usage by defining a session as the difference between the start times of two logged events. The authors allocated the time-windows *intuitively*, but they further corroborate their choice by conducting a brief survey with 10 individuals who were asked to divide the working day into four major time segments, which resulted in the following division: 05:00 to 12:00, 12:00 to 18:00, 18:00 to 21:00, and 21:00 to 05:00 (Table 3). The whole dataset was used to conduct the experiments: all logs and the same observation period (the whole semester).

Table 3. Time-Windows

	Morning	**Afternoon**	**Evening**	**Night**
Sher et al. (2022)	05:00-12:00	12:00-18:00	18:00-21:00	21:00-05:00
Case study	05:00-13:00	13:00-20:00	20:00-01:00	01:00-05:00

Session identification settings: The focus of interest is the broad aim of using the platform in general. Thus we select the "study session" type, since it allows us to capture all platform-related activities, and the "task," since this grouped granularity would provide more valuable insight into students' interactions (Definition 7). We propose to use both session identification methods, authentication logs and STT (Section 6.1), because they complement each other and they adhere to the experiment mentioned in the reference paper (Sher et al., 2022). For the STT, we suggest using the temporal metric of time-off-task (*estimated inactivity duration*—Section 5.1) and setting it to 3.89 minutes, according to the recommendation (Section 5.2). In addition, because we already know that the platform is used for certain required functions due to our domain knowledge, we eliminate attendance-only sessions (Section 6.1) because they provide no additional information about student behaviour.

Time-window identification: Looking at Figure 6, we detect numerous outliers on the time-window identification plot. The average time of the whiskers is 51 min 15 s, and the average time of the box elements, which are all contained within the same hour, is 28 min 30 s.

Analyzing the night and morning hours (Figure 7a), two horizontal lines do not intersect any of the boxplot elements: 05:00 and 06:00. To determine the optimal split-point, we compare the whisker and the median differences between the preceding and subsequent boxplots and select the one with the most significant difference. A higher value indicates that two boxplots are further apart. In particular, the upper whisker of the 04:00 boxplot extends up to 04:56, and the lower whisker of the 05:00 boxplot begins at 05:09, which means a 13 min difference. The upper whisker of the 05:00 boxplot extends up to 05:59, and

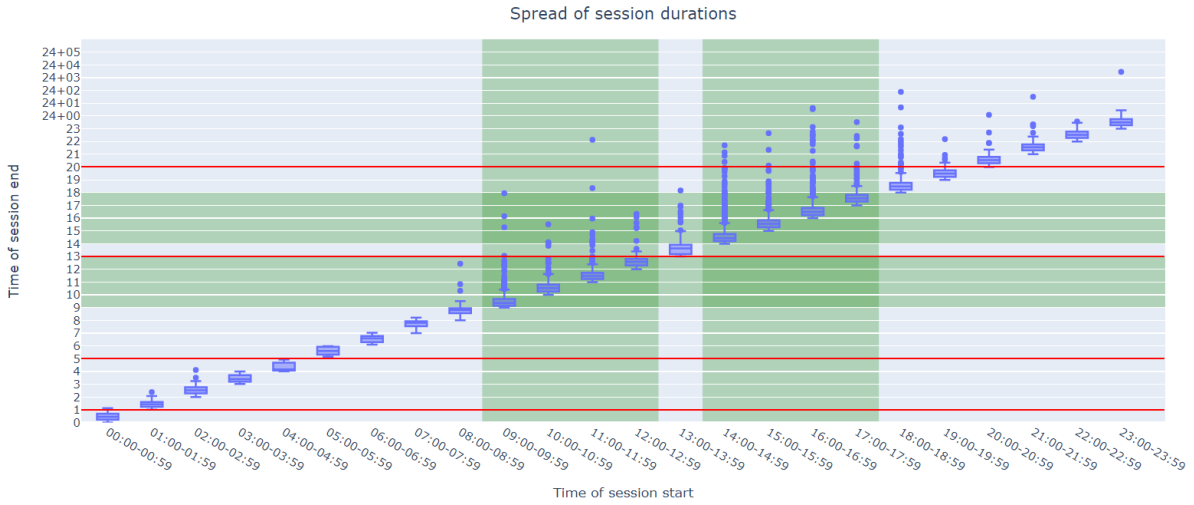
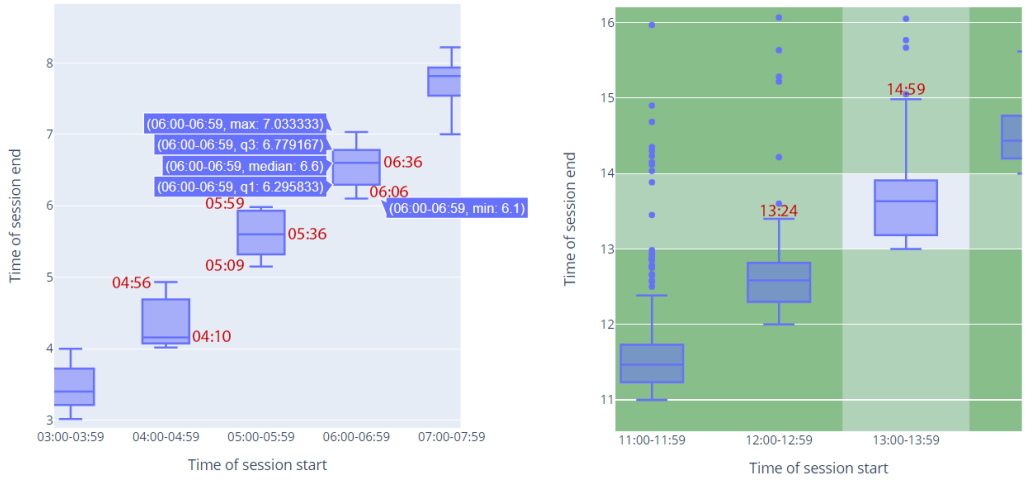


Figure 6. Time-window identification plot for the case study. It depicts the plot obtained by replicating the experiment of Sher and colleagues (2022) on our dataset. The red horizontal lines represent the split-points obtained by using the proposed approach.

the lower whisker of the 06:00 boxplot begins at 06:06, which means a 7 min difference. Moreover, the median difference between the 04:00 and 05:00 boxplots is 1 h 26 min, whereas between the 05:00 and 06:00 boxplots it is 1 h. Therefore, the 05:00 split-point is chosen to represent the night-morning division.



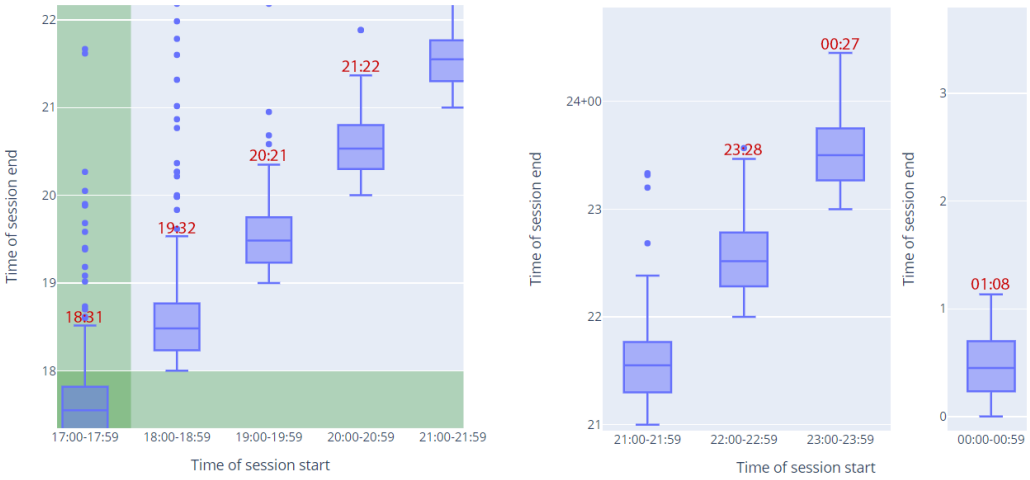
(a) Night/morning split **(b)** Morning/afternoon split

Figure 7. Case study. Time-window identification plot in detail (first half of the day.)

From our domain knowledge (Section 3), we know that the optimal time to divide the morning and afternoon activities would be either 13:00 (the conclusion of the morning classes) or 14:00 (the beginning of the afternoon laboratories). However, based on the plot in Figure 6, in both cases, outliers are intersected. However, as illustrated in Figure 7b, the 13:00 horizontal line intersects the upper whisker of the 12:00 boxplot, which extends up to 13:24. In contrast, the 14:00 horizontal line intersects the upper whisker of the 13:00 boxplot, which extends up to 14:59. Thus, we choose the split-point at 13:00 to represent the morning-afternoon time-window division.

The ideal split-point between afternoon and evening activities (Figure 8a), between 18:00 and 21:00 hours, is 20:00. Also in these cases, the horizontal line always intersects outliers, but the 20:00 split-point provides the smaller previous-hour whisker cut (21 min compared to 31 min, 32 min, and 22 min) and allows the majority of afternoon boxplot outliers (beginning with the 13:00 split-point) to be contained within the afternoon time-window.

The final split pertains to the evening-night divide (Figure 8b), where it can be observed that from 23:00 to 01:00, in



(a) *Afternoon/evening split* **(b)** *Evening/night split*

Figure 8. *Case study. Time-window identification plot in detail (second half of the day).*

addition to the intersection of outliers, the 01:00 split-point represents the smaller previous-hour whisker cut. Finally, we derive four time-windows: 05:00–13:00, 13:00–20:00, 20:00–01:00, and 01:00–05:00, as shown in Table 3.

We can observe students' tendency to continue their afternoon activities after 18:00. Figure 6 shows that a split-point at 18:00 would have divided many sessions into two time-windows. Conversely, if we had postponed the evening split-point to 20:00 since the data indicate that students tend to work until 20:00, a limited number of sessions would have been divided, i.e., a more significant number of sessions would have been contained in the same time-window. Consequently, because the core premise of our approach is to reduce the likelihood that a session is split into two different time-windows, our time-window segmentation enables us to analyze interconnected events that are at the basis of behavioural regularities together.

Our definition of the evening (20:00–01:00) differs from the work of Sher and colleagues (2022) (18:00–21:00), which has been intuitively set and confirmed by a survey to more closely fit with how students might organize their day. Nevertheless, the reference does not specify which time slots were chosen before the survey. Hence, we can only observe that our data-driven approach, which uses the study session definition (Section 5.1), offers valuable suggestions to move the split-point to 20:00, highlighting the fact that the potential intuitively selected 18:00 split-point would not have optimally suited our data.

7. Conclusion and Future Work

In this paper, we have presented a study aimed at supporting the identification of session-based time-windows. This work is not intended to examine student behaviour, student engagement, or student participation in online classes, but to help any analysis concerning students' temporal behaviour when learning online. To accomplish our goal of analyzing records to identify time-windows during online learning, we only require logs of online student interaction. Clearly, students can perform activities outside the platform while learning, but we cannot capture them and they are outside the scope of our analysis. In contrast to prior work, our method does not rely on an arbitrary threshold to identify sessions; instead, it is based on algorithms that assist us in calculating the STT. We first provided a clear and well-defined method for identifying three distinct session types for the education environment, *study*, *course*, and *learning* session, with data-driven algorithms based on the duration of inactivity during online learning. Then, using data-driven methodologies, we defined a process for supporting the investigator in identifying time-windows based on the detected sessions. To determine the optimal parameter settings for session identification and locating suitable time-windows, we developed a visual tool that can be accessed online for hands-on practice.

Our experiments revealed that time-windows vary based on many factors, such as the session type, the observation period, the course selection, and the given dataset, showing the influence of these factors on student temporal behaviour. This demonstrates that it is impossible to identify unique time-windows that are optimal in every circumstance and that our data-driven method yields outcomes which differ from any intuitively assigned time-windows.

We would like to stress that our methodology provides (i) some flexibility to the investigator in selecting a number of options according to the real-world scenario under analysis (Section 6.1) and (ii) visual information about both data evidence and domain knowledge, enabling the investigator to determine time-windows with complete awareness and flexibility.

The methodology is only a preliminary step for identifying time-windows to be used in temporal learning pattern analysis. Once the time slots have been identified, it will be possible to use them to determine, for instance, whether the action or task that

occurs in that specific time slot could represent a temporal pattern. Therefore, this work will proceed in two parallel directions in the future. On the one hand, we will conduct surveys to elicit students' explicit session activities to assess the quality of the identified session-based time-windows. Moreover, we will analyze our tool with the involvement of potential users to evaluate its effectiveness. To conduct a statistically valid study, the evaluation of the effectiveness requires (1) an experimental design able to identify the possible types of users of the tool; (2) the definition of the strategy to recruit the users; (3) the presentation of the user study for an ethics committee; and (4) the comparison of the organization of a course after analyzing the temporal behaviour of the students *with* and *without* the support of our visual tool, for assessing the validity of using the tool through bi-testing. On the other hand, the identified time-windows will be used to support temporal student learning pattern analyses and to evaluate how they affect the outcomes.

The proposed approach for estimating the STT currently takes into consideration the activities of all students of the platform for the study session or all students enrolled in a specific course for the course session. Although this is suitable for a collective understanding, it is hard to perform personalized analyses for each student. Thus, to this end, it would be interesting to explore the possibility of designing a method for estimating personalized STT adequate to the specific student's habits.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work has been partially supported by EU—Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019—Integrating Activities for Advanced Communities," Grant Agreement n.871042; "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu); the scheme "HORIZON-INFRA-2021-DEV-02—Developing and consolidating the European research infrastructures landscape, maintaining global leadership (2021)," Grant Agreement n.101079043; "SoBigData RI PPP: SoBigData RI Preparatory Phase Project" by NextGenerationEU—National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR)—Project: "SoBigData.it—Strengthening the Italian RI for Social Mining and Big Data Analytics"—Prot. IR000001 3—Avviso n. 3264 del 28/12/2021; and by PNRR—M4C2—Investimento 1.3, Partenariato Esteso PE00000013—"FAIR—Future Artificial Intelligence Research"—Spoke 1 "Human-Centered AI" funded by the European Commission under the NextGeneration EU programme.

References

Alvarez Sainz, M., Ferrero, A. M., & Ugidos, A. (2019). Time management: Skills to learn and put into practice. *Education + Training*, *61*(5), 635–648. https://doi.org/10.1108/ET-01-2018-0027

Asadianfam, S., & Mohammadi, M. (2014). Identify navigational patterns of web users. *International Journal of Computer-Aided Technologies (IJCAx)*, *1*(1), 398–426. https://airccse.org/journal/ijcax/papers/1114ijcax01.pdf

Ba-Omar, H., Petrounias, I., & Anwar, F. (2007). A framework for using web usage mining to personalise e-learning. In *Proceedings of the Seventh IEEE International Conference on Advanced Learning Technologies* (ICALT 2007), 18–20 July 2007, Niigata, Japan (pp. 937–938). IEEE. https://doi.org/10.1109/ICALT.2007.13

Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A., & Vigna, S. (2008). The query-flow graph: Model and applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (CIKM 2008), 26–30 October 2008, Napa Valley, California, USA (pp. 609–618). ACM. https://doi.org/10.1145/1458082.1458163

Britzolakis, A. (2020). *Design and development of a web-based data visualization software for political tendency identification of Twitter's users using Python Dash Framework.* [Master's thesis, Hellenic Mediterranean University, School of Engineering (ScENG) MSc in Informatics & Multimedia]. https://apothesis.lib.hmu.gr/bitstream/handle/20.500.12688/9510/BritzolakisAlexandros2020.pdf

Brooks, C. A., Thompson, C., & Teasley, S. D. (2015). A time series interaction analysis method for building predictive models of learners using log data. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (LAK 2015), 16–20 March 2015, Poughkeepsie, New York, USA (pp. 126–135). ACM. https://doi.org/10.1145/2723576.2723581

Carroll, J. (1963). A model of school learning. *Teachers College Record*, *64*(8), 1–9. https://doi.org/10.1177/016146816306400801

Catledge, L. D., & Pitkow, J. E. (1995). Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems*, *27*(6), 1065–1073. https://doi.org/10.1016/0169-7552(95)00043-7

Chen, B., Knight, S., & Wise, A. (2018). Critical issues in designing and implementing temporal analytics. *Journal of Learning Analytics*, *5*(1), 1–9. https://doi.org/10.18608/jla.2018.53.1

Chitraa, V., & Davamani, A. S. (2010). A survey on preprocessing methods for web usage data. *International Journal of Computer Science and Information Security*, *7*(3), 78–83. https://doi.org/10.48550/arXiv.1004.1257

Chiu, M. M., Molenaar, I., Chen, G., Wise, A. F., & Fujita, N. (2014). Micro-analysis of collaborative processes that facilitate productive online discussions: Statistical discourse analyses in three studies. In E. Barbera & P. Reimann (Eds.), *Assessment and evaluation of time factors in online teaching and learning* (pp. 232–263). IGI Global. https://doi.org/10.4018/978-1-4666-4651-3.ch009

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the World Wide Web. In *Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence,* 3–8 November 1997, Newport Beach, California, USA (pp. 558–567). IEEE. https://doi.org/10.1109/TAI.1997.632303

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining World Wide Web browsing patterns. *Knowledge and Information Systems*, *1*(1), 5–32. https://doi.org/10.1007/bf03325089

Credé, M., & Kuncel, N. R. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. *Perspectives on Psychological Science*, *3*(6), 425–453. https://doi.org/10.1111/j.1745-6924.2008.00089.x

Du, X., Zhang, M., Shelton, B. E., & Hung, J.-L. (2019). Learning anytime, anywhere: A spatio-temporal analysis for online learning. *Interactive Learning Environments*, *30*(1), 34–48. https://doi.org/10.1080/10494820.2019.1633546

Fatima, B., Ramzan, H., & Asghar, S. (2016). Session identification techniques used in web usage mining: A systematic mapping of scholarly literature. *Online Information Review*, *40*(7), 1033–1053. https://doi.org/10.1108/OIR-08-2015-0274

Fincham, E., Whitelock-Wainwright, A., Kovanović, V., Joksimović, S., van Staalduinen, J.-P., & Gašević, D. (2019). Counting clicks is not enough: Validating a theorized model of engagement in learning analytics. In *Proceedings of the Ninth International Conference on Learning Analytics and Knowledge* (LAK 2019), 4–8 March 2019, Tempe, Arizona, USA (pp. 501–510). ACM. https://doi.org/10.1145/3303772.3303775

Gitinabard, N., Barnes, T., Heckman, S., & Lynch, C. F. (2019). What will you do next? A sequence analysis on the student transitions between online platforms in blended courses. In *Proceedings of the 12th International Conference on Educational Data Mining* (EDM 2019), 2–5 July 2019, Montréal, Québec, Canada (pp. 59–68). https://doi.org/10.48550/arXiv.1905.00928

Goda, Y., Yamada, M., Kato, H., Matsuda, T., Saito, Y., & Miyagawa, H. (2015). Procrastination and other learning behavioral types in e-learning and their relationship with learning outcomes. *Learning and Individual Differences*, *37*, 72–80. https://doi.org/10.1016/j.lindif.2014.11.001

Goodyear, P. (2005). Educational design and networked learning: Patterns, pattern languages and design practice. *Australasian Journal of Educational Technology*, *21*(1). https://doi.org/10.14742/ajet.1344

Goodyear, P., & Retalis, S. (2010). *Technology-enhanced learning: Design patterns and pattern languages* (P. Goodyear & S. Retalis, Eds.). Brill. https://doi.org/10.1163/9789460910623

Heraud, J.-M., France, L., & Mille, A. (2004). Pixed: An ITS that guides students with the help of learners' interaction log. In *Workshop Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes: Seventh International Conference on Intelligent Tutoring Systems* (ITS 2004), 30 August–3 September 2004, Maceió, Alagoas, Brazil (pp. 57–64).

Heyer, S., Oberhuemer, P., Zander, S., & Prenner, P. (2007). Making sense of IMS learning design level B: From specification to intuitive modeling software. In *Creating New Learning Experiences on a Global Scale: Proceedings of the Second European Conference on Technology Enhanced Learning* (EC-TEL 2007), 17–20 September 2007, Crete, Greece (pp. 86–100). Springer. https://doi.org/10.1007/978-3-540-75195-3_7

Hossain, S., Calloway, C., Lippa, D., Niederhut, D., & Shupe, D. (2019). Visualization of bioinformatics data with Dash Bio. In *Proceedings of the 18th Python in Science Conference* (SciPy 2019), 8–14 July 2019, Austin, Texas, USA. https://doi.org/10.25080/Majora-7ddc1dd1-012

IMS Global Learning Consortium and others. (2003). IMS Learning Design Information Model. Ver. 1.0 Final Specification. https://www.imsglobal.org/learningdesign/ldv1p0/imsld%5C_infov1p0.html

Jafari, M., SoleymaniSabzchi, F., & Jamali, S. (2013). Extracting users' navigational behavior from web log data: A survey. *Journal of Computer Sciences and Applications*, *1*(3), 39–45. https://doi.org/10.12691/jcsa-1-3-3

Jansen, B. J., Spink, A., Blakely, C., & Koshman, S. (2007). Defining a session on web search engines. *Journal of the American Society for Information Science and Technology*, *58*(6), 862–871. https://doi.org/10.1002/asi.20564

Jones, R., & Klinkner, K. L. (2008). Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (CIKM 2008), 26–30 October 2008, Napa Valley, California, USA (pp. 699–708). ACM. https://doi.org/10.1145/1458082.1458176

Jovanovic, J., Gasevic, 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. https://doi.org/10.1016/j.iheduc.2017.02.001

Khan, I., & Pardo, A. (2016). Data2U: Scalable real time student feedback in active learning environments. In *Proceedings of the Sixth International Conference on Learning Analytics and Knowledge* (LAK 2016), 25–29 April 2016, Edinburgh, UK (pp. 249–253). ACM. https://doi.org/10.1145/2883851.2883911

Kovanović, V., Gašević, D., Dawson, S., Joksimovic, S., & Baker, R. (2016). Does time-on-task estimation matter? Implications on validity of learning analytics findings. *Journal of Learning Analytics*, *2*(3), 81–110. https://doi.org/10.18608/jla.2015.23.6

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* (LAK 2015), 16–20 March 2015, Poughkeepsie, New York, USA (pp. 184–193). ACM. https://doi.org/10.1145/2723576.2723623

Lee, S., Kim, K. S., Shin, J., & Park, J. (2021). Tracing knowledge for tracing dropouts: Multi-task training for study session dropout prediction. In I.-H. Hsiao, S. Sahebi, F. Bouchet, & J.-J. Vie (Eds.), *Proceedings of the 14th International Conference on Educational Data Mining* (EDM 2021), 29 June–2 July 2021, Paris, France (pp. 641–647). https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_174.pdf

Lee, Y. (2018). Effect of uninterrupted time-on-task on students' success in Massive Open Online Courses (MOOCs). *Computers in Human Behavior*, *86*, 174–180. https://doi.org/10.1016/j.chb.2018.04.043

Lee, Y., Shin, D., Loh, H., Lee, J., Chae, P., Cho, J., Park, S., Lee, J., Baek, J., Kim, B., & Choi, Y. (2020). Deep attentive study session dropout prediction in mobile learning environment. In *Proceedings of the 12th International Conference on Computer Supported Education—Volume 1: CSEDU,* 2–4 May 2020, online (pp. 26–35). https://doi.org/10.5220/0009347700260035

Lokeshkumar, R., Sindhuja, R., & Sengottuvelan, P. (2014). A survey on preprocessing of web log file in web usage mining to improve the quality of data. *International Journal of Emerging Technology and Advanced Engineering*, *4*, 229–234. https://www.ijetae.com/files/Volume4Issue8/IJETAE_0814_35.pdf

Lucchese, C., Orlando, S., Perego, R., Silvestri, F., & Tolomei, G. (2013). Discovering tasks from search engine query logs. *ACM Transactions on Information Systems (TOIS)*, *31*(3), 1–43. https://doi.org/10.1145/2493175.2493179

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., & Delgado-Kloos, C. (2018). Predicting learners' success in a self-paced MOOC through sequence patterns of self-regulated learning. In V. Pammer-Schindler, M. Pérez-Sanagustín, H. Drachsler, R. Elferink, & M. Scheffel (Eds.), *Proceedings of the 13th European Conference on Technology Enhanced Learning* (EC-TEL 2018), 3–5 September 2018, Leeds, UK (pp. 355–369). Springer. https://doi.org/10.1007/978-3-319-98572-5_27

Marquardt, C. G., Becker, K., & Ruiz, D. D. A. (2004). A pre-processing tool for web usage mining in the distance education domain. In J. Bernardino & B. C. Desai (Eds.), *Proceedings of the International Database Engineering and Applications Symposium, 2004* (IDEAS 2004), 7–9 July 2004, Coimbra, Portugal (pp. 78–87). IEEE. https://doi.org/10.1109/IDEAS.2004.1319780

Matcha, W., Gašević, D., Jovanović, J., Uzir, N. A., Oliver, C. W., Murray, A., & Gasevic, D. (2020). Analytics of learning strategies: The association with the personality traits. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge* (LAK 2020), 23–27 March 2020, Frankfurt, Germany (pp. 151–160). ACM. https://doi.org/10.1145/3375462.3375534

Molenaar, I., & Friend Wise, A. (2022). Temporal aspects of learning analytics—Grounding analyses in concepts of time. In C. Lang, G. Siemens, A. Friend Wise, D. Gašević, & A. Merceron (Eds.), *The handbook of learning analytics* (pp. 66–76). SOLAR. https://doi.org/10.18608/hla22.007

Munk, M., & Drlík, M. (2011). Impact of different pre-processing tasks on effective identification of users' behavioral patterns in web-based educational system. *Procedia Computer Science*, *4*, 1640–1649. https://doi.org/10.1016/j.procs.2011.04.177

Nadjarbashi-Noghani, M., & Ghorbani, A. A. (2004). Improving the referrer-based web log session reconstruction. In *Proceedings of the Second Annual Conference on Communication Networks and Services Research* (CNSR 2004), 21 May 2004, Fredericton, New Brunswick, Canada (pp. 286–292). IEEE. https://doi.org/10.1109/DNSR.2004.1344740

Nguyen, Q. (2020). Rethinking time-on-task estimation with outlier detection accounting for individual, time, and task differences. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge* (LAK 2020), 23–27 March 2020, Frankfurt, Germany (pp. 376–381). ACM. https://doi.org/10.1145/3375462.3375538

Panzarasa, P., Kujawski, B., Hammond, E. J., & Roberts, C. M. (2015). Temporal patterns and dynamics of e-learning usage in medical education. *Educational Technology Research and Development*, *64*(1), 13–35. https://doi.org/10.1007/s11423-015-9407-4

Riel, J., Lawless, K. A., & Brown, S. W. (2018). Timing matters: Approaches for measuring and visualizing behaviours of timing and spacing of work in self-paced online teacher professional development courses. *Journal of Learning Analytics*, *5*(1), 25–40. https://doi.org/10.18608/jla.2018.51.3

Romero, C., Romero, J. R., & Ventura, S. (2014). A survey on pre-processing educational data. In A. Peña-Ayala (Ed.), *Educational data mining: Applications and trends* (pp. 29–64). Springer. https://doi.org/10.1007/978-3-319-02738-8_2

Rotelli, D., & Monreale, A. (2022). Time-on-task estimation by data-driven outlier detection based on learning activities. In *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (LAK 2022), 21–25 March 2022, online (pp. 336–346). ACM. https://doi.org/10.1145/3506860.3506913

Schmitz, B., & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary Educational Psychology*, *31*(1), 64–96. https://doi.org/10.1016/j.cedpsych.2005.02.002

Sher, V., Hatala, M., & Gašević, D. (2022). When do learners study? An analysis of the time-of-day and weekday-weekend usage patterns of learning management systems from mobile and computers in blended learning. *Journal of Learning Analytics*, *9*(2), 1–23. https://doi.org/10.18608/jla.2022.6697

Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *INFORMS Journal on Computing*, *15*(2), 171–190. https://doi.org/10.1287/ijoc.15.2.171.14445

Thindwa, H. (2016). The role of technology in improving quality of teaching in higher education: An international perspective. In Information Resources Management Association (Ed.), *Teacher education: Concepts, methodologies, tools, and applications* (pp. 207–227). IGI Global. https://doi.org/10.4018/978-1-5225-0164-0.ch012

Wang, H., Tlili, A., Lämsä, J., Cai, Z., Zhong, X., & Huang, R. (2023). Temporal perspective on the gender-related differences in online learning behaviour. *Behaviour & Information Technology*, *42*(6), 671–685. https://doi.org/10.1080/0144929X.2022.2039769

Weinstein, C. E., Palmer, D. R., & Schultz, A. (2002). *Lassi user's manual*. H&H Publishing Company. https://www.hhpublishing.com/LASSImanual.pdf