

Process Mining for Analyzing Open Questions Computer-Aided Examinations

Saimir Bala^[0000-0001-7179-1901], Kate Revoredo^[0000-0001-8914-9132], and Jan
Mendling^[0000-0002-7260-524X] *

Humboldt University of Berlin, Berlin, Germany
{saimir.bala,kate.revoredo,jan.mendling}@hu-berlin.de

Abstract. Computer-based education relies on information systems to support teaching and learning processes. These systems store trace data about the interaction of the learners with their different functionalities. Process mining techniques have been used to evaluate these traces and provide insights to instructors on the behavior of students. However, an analysis of students behavior on solving open-questioned examinations combined with the marks they received is still missing. This analysis can support the instructors not only on improving the design of future edition of the course, but also on improving the structure of online and physical evaluations. In this paper, we use process mining techniques to evaluate the behavioral patterns of students solving computer-based open-ended exams and their correlation with the grades. Our results show patterns of behavior associated to the marks received. We discuss how these results may support the instructor on elaborating future open question examinations.

Keywords: education, process mining, educational process mining, exam process

1 Introduction

Educational process mining [1] analyzes data generated from educational systems using process mining techniques. These analyses may support the course instructors for example on understanding how students engage with self-assessment [2] or how the students behave while using the online educational systems [3]. Typically, these kind of analyses focus on online courses, such as the ones provided by Massive Open Online Course (MOOC) platforms like Coursera¹, Edx², etc. In these settings the courses are designed for being taught online. Consecutively also students assessments is performed online using closed-answer questions [3,4].

However, with the advent of the COVID-19 pandemic, a new reality emerged: courses that were designed to be held in presence had to switch to online mode

* This work was supported by the Einstein Foundation Berlin [grant number EPP-2019-524, 2022]

¹ <https://www.coursera.org>

² <https://www.edx.org>

due to the various lockdowns. Oftentimes this transfer from physical mode to online mode had to be performed with short notice, leaving the instructors little or not time to design the course anew. As a result, many courses “switched” to online by simply mimicking their in-presence version. Thus, previously planned on-paper exams, were simply replaced by online documents to be downloaded by the students, performed within a remotely-controlled environment (i.e., monitoring students via webcam, microphone, screenshots) and uploaded again to the system [5]. These kind of exams are referred to as *open questions computer-aided examinations*.

In order to make it possible for the teacher or other authorities to check for students misconduct during the exam at a later stage, the monitoring data are usually persisted in event logs as *trace data* [6]. In this way, such setting opens up to a unique opportunity to use process mining to gain further insights on the *exam-taking process*. Specifically, mining techniques can be used to support the instructors on understanding how the students behave when solving the exam, for example to understand which question was more or less demanding. Also, confronting the behavior for solving the exam with the marks the students received may provide other insights to the instructors. For example, if certain behavior leads to better performance, if the most demanding question was also the one with lower marks meaning that the students did not acquire the knowledge.

In this paper, we use a multi-method approach based on process mining [7] to analyze the trace data generated from the interaction of students with an online system while doing an exam with open questions. We used data from two master course exams. We enriched the trace data with the marks the students received for each of the questions. The results show that there is a pattern on solving the exam when considering the topmost performers students. Also, there is a relation between the time spent by the students on solving the question with the marks they received for the question. With this research we contribute to the area for educational process mining by showing how process mining can be used also to support courses designed as physical.

The rest of the paper is structured as follows. Section 2 discusses the related work. Section 3 presents our method for evaluating students behavior on solving an online exam. Section 4 describes our scenario of application including the setting, results and some discussions. Section 5 concludes our work and provide some future directions.

2 Related Work

Educational data has been exploited by process mining techniques for various analyses [8] and in several ways. In [3] the data is grouped considering the grades and the behavior of the students while using an online educational tool. Process mining techniques are then applied to the different groups showing that the models discovered are more comprehensible and with higher fitness than the models learned using the whole data. In [9] data from a Massive Open Online

Course (MOOC) was used to analyze the behavior of the students during the learning phase confronted with the final marks they achieve in the course.

Our literature review did not find many works related to ours, i.e., that focus on using process mining for analyzing the behavior of the students considering an *online assessment*. In the remaining of this section we outline the contributions of works who take online assessment into account.

In [2] process mining is used to evaluate how students engage with self-assessment and formative assessment and how these two types of assessments are related. In [4] process mining is used to evaluate the navigation behavior of students when answering to a close-ended online test. Also, a navigation reference model is used for conformance checking. The results of the paper show that the navigation behavior impacts on the performance of the students. In [10] a system to automatically evaluate the performance of students was proposed. Process discovery is used to learn the process used by the students when doing an online test that requires the use of a special software, e.g., ERP system. The process discovered represent how the student behaved to achieve the given business scenario. Based on the learned model students' performance is automatically evaluated. In [11] different techniques of process mining were used to evaluated assessment data from online multiple choice tests. Data from two exams were considered where in a first study the questions must be answered in a strict order and immediately after the students could receive feedback and learn the correct answer. In the second study, the student could choose the order for answering the questions and they could revisit earlier questions and their answer. In [12] process mining was used to examine self-regulated learning (SRL) of students. By analyzing data recorded during think-aloud sessions, differences were found in the frequency and order of the performed tasks. In [13] a framework called Trace-SRL was proposed to analyze SRL processes, using stochastic process mining under the theoretical lens of SRL. Findings include the discovery of different learner-strategy types along with specific patterns.

While related, none of the above-mentioned works focuses on evaluating *computer-aided examinations with open-ended questions*. In the following section, we describe how this format of examination can be evaluated. Also, we compare the behavior of the students when answering the test with the marks they received for the questions.

3 Method

Computer-aided examinations with open-ended questions belong to the context of computer-aided teaching and learning, In this environment there are two main actors, the instructor and the student. The instructor is responsible for *designing* the course, *teaching* and *grading* the students. The students are responsible for *learning* the content taught to them and perform an *examination* that assesses how well they acquired the content taught in the course. During the course design phase the instructor designs the classes to be taught and also how the examination will be. In this environment, the examination is performed using

an online system that is able to store the interactions of the students with the system during the examination. Also, the grading of the examination is made available digitally.

Our method aims at improving the instructors knowledge about the exam-process. To this end, its input is constituted by the generated trace data and the grades of the students. Next, it encodes them onto an enriched event log. Process mining techniques and statistical methods are used to analyze this enriched event log, presenting the discovered knowledge to the instructor. Based on this knowledge the instructor may change the design of future editions of the course, which may include more teaching time on concepts that were shown to be not clear to the students or changing the order of the concepts being evaluated (i.e., questions) in the examination. Figure 1 depicts a sketch of our method. The shaded steps are the steps of the computer-aided teaching and learning process on which this research focuses.

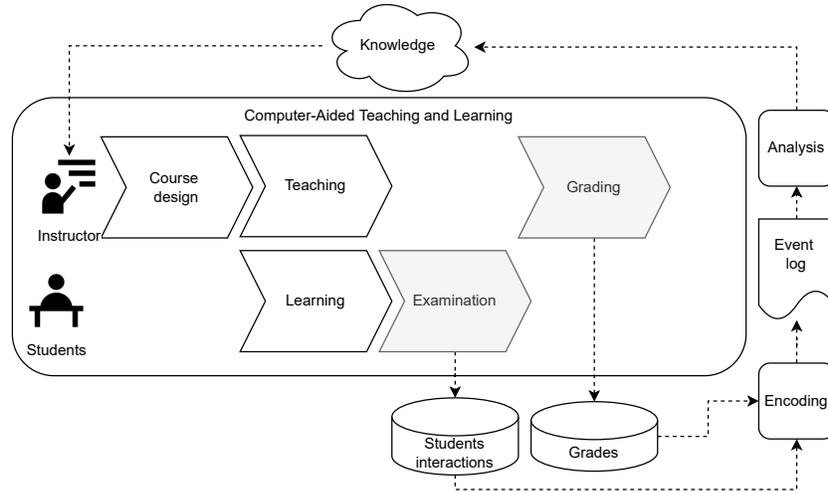


Fig. 1: Exploiting process mining to analyze exams in order to gather knowledge for improving computer-aided teaching & learning processes.

The encoding of the event log is done in two steps. The first step creates a standard event log in the XES [14] format with case ID, activity and timestamp. And the second enriches the created event log with attributes that correspond to the grades received in the questions and in the overall exam. Also, attributes that store the duration of working on each question are included in the event log.

For generating the event log the method requires that the online system being used stores data that allow one to identify when the student was working in each of the questions. Given that the goal is to analyze the student behavior on

performing an online exam, the student matriculation number defines a case in the event log and the student interactions with the system define the activities.

We focus on collecting insights on how the students answered the questions of the exam, therefore we created four activity templates: *Look Question X*, *Work on Question X*, *Save Draft of Question X* and *Submit Question X*, where X is a placeholder for the question number. The first activity starts when the student opens question X and finishes when the student either closes question X, starts working on question X or save or submit question X. The second activity starts when the student starts writing on the environment and finishes when the student closes the question or presses the save or submit button in the system. The third activity happens when the student pushes the button to save a draft of the question. The fourth activity happens when the student pushes the button to submit the question. Saving a draft or submitting a question do not have a meaningful duration. The duration that we are interested on is the duration of the *Work on Question X* activity, which represents the overall time that the student took to answer the question. The structure of the final traces in the event log is $Trace_i = \langle Student_i, Activity, Time, Grade_{Q_1}, \dots, Grade_{Q_n}, Duration_{Q_1}, \dots, Duration_{Q_n} \rangle$.

For the analysis phase process mining techniques [7] are used. Process discovery is used to explore the sequence behavior of solving the exam searching for possible patterns. Process data such as duration of the activities, activities most frequently executed were collected for the analysis. The choice of the process data collected is of the instructor depending on the investigation he or she wants to do.

4 Application

As an application scenario, we applied our method on a case of open question examinations, henceforth called *Exam1* and *Exam2*. These examinations stemmed from two master courses of an Austrian university. The two courses were designed for in-presence teaching and examinations. However, due to a COVID-19 wave the classes and the exams were moved to online.

For what concerns the exams, all the setup was kept the same as to a written exam taken physically. All the questions were made available online in the Teaching and Learning information system in the same way that they were available in the paper format. The students could visit the questions in the order they wanted and as many times as they wanted. It was possible to save a draft of the answers until the submission of the final answers. Each exam had to be performed in 90 minutes, however the students were given 10 minutes more than planned for the physical exam to compensate any possible infrastructure issue, summing up a maximum of 100 minutes for doing the exam.

In the following, we describe the details of applying our method to analyze the behavior of the students when answering the questions of *Exam1* and *Exam2* and how this behavior is related to the grades they achieved. Section 4.1 describes the event logs generated from the data logged by the Moodle system and the grades achieved by the students. Section 4.2 presents the results found. And, Section 4.3 discusses our findings.

4.1 Setting

Exam1 had 8 questions and *Exam2* had 4 questions. Thus, the event log for each of the exams were composed by 32 and 16 distinct activities, respectively. The teaching and learning system stored every interaction event along with a timestamp. This timestamp was used to set when each activity started. The students matriculation identifiers defined the cases in the event log. For *Exam1* 61 students completed the exam and for *Exam2* 27 students. Thus, the event logs had 61 and 27 cases, respectively.

The event log was enriched with the grade received for each of the questions and also the final grade on the exam. Both exams have a maximum of 100 points. For *Exam1* the points were split equally, i.e., each of the questions had a maximum of 12.5 points. For *Exam2*, three questions (1,2 and 4) had maximum of 20 each and one question (3) had a maximum of 40 points. The event logs were then filtered by the achieved final grade generating two event logs for each exam. One event log was composed by the traces of the students that achieved more than 80 (inclusive) in the exam, i.e., the topmost performers (*TP*) of the exam. The other event log was composed by the traces of the students that achieved less than 50 (inclusive), i.e., the lowermost performers (*LP*) of the exam. Table 1 provides details on the event logs generated. For analyzing the data we used the

Table 1: Event logs description

Event Log	Cases	Events	Activities	Median case duration	Mean case duration	Min. activity frequency	Median act. freq.	Mean act. freq.	Max act. frequency	Act. Frequency std. dev.
<i>Exam1</i>	61	7173	32	88.9 mins	86.7 mins	2	178	224.16	769	199.15
<i>Exam1_{TP}</i>	24	2708	31	89.4 mins	87.6 mins	1	76	87.35	349	84.85
<i>Exam1_{LP}</i>	7	820	29	87.8 mins	82.5 mins	1	27	28.28	85	23.26
<i>Exam2</i>	27	2892	16	99.3 mins	91 mins	2	115	180.75	707	203.53
<i>Exam2_{TP}</i>	7	698	15	84.1 mins	84.4 mins	1	27	46.53	163	50.58
<i>Exam2_{LP}</i>	10	1144	16	103.8 mins	97.2 mins	1	52	71.5	274	75.2

Disco³ tool for process mining and R⁴ software for correlation analysis and plot generation. The analyses were guided by the following questions:

Q1: What are patterns in the behavior of exam solving?

³ <https://fluxicon.com/disco>

⁴ <https://www.r-project.org>

Q2: How does the grade correlate with the time spent to solve a question?

4.2 Results

The number of variants is the same as the number of cases, amounting to 61 variants for *Exam1* and 27 variants for *Exam2*, which shows that every student used a different strategy to solve the exam.

Figure 2 and Figure 3 depict the process models learned from the event logs generated for *Exam1* and *Exam2*, respectively. The processes on the left (a) were learned from *Exam1_{TP}* and *Exam2_{TP}* event logs, while the processes on the right (b) were learned from *Exam1_{LP}* and *Exam2_{LP}* event logs.

The activities most executed in all cases are the *Work on Question X*, given that these activities are executed when the students are working on the answers for question X. By analyzing the difference in the color of these activities it is possible to notice which question the students took more time for answering. There are self cycle in some of these activities, which means that the teaching and learning information system automatically saved the draft of the answer. This happens when the student is continually changing the content of the question, so periodically the system auto saves the content. Cycles are observed in all processes, meaning that the students did not work in one question and submitted it. They chose to save the draft of the answers and they returned to the question either to change its content or to validate before submitting it. Also, the control flow analysis of the processes show that in general the topmost performer students solved the exam following the order presented while the lowermost performer students solved the exam in a more chaotic way.

Figure 4 depicts the distribution of points for each question. For *Exam1* the majority of the students achieved the maximum points for questions 2, 6 and 8, given the median close to 12.5. Question 7 has a diversity of points, which indicate that it was the most controversial question in this exam. It is potentially a question about concepts not well understood by the students and thus a concept more deeply discussed with the students in future editions of the course. *Exam2* seems a more hard exam given that the majority of the students did not score the maximum points of the questions. Question 3 is the question with more variation on the points received by the students. It seems the hardest question in this exam, which it is expected given that it is the only question with maximum of 40 points. Some of the outliers presented in *Exam1* correspond to students in the lowermost performers students. The rest correspond to students that achieve between 50 and 80 points.

Figure 5 depicts the distribution of the time spent to solve each of the questions. The duration considers only the time spent on working on a particular question. It is expressed in seconds. For *Exam1* the question that was done faster by the students were Question 6, while the question that demanded more time was question 5. When confronting this result with the one presented in Figure 4 it seems that Question 6 was the easiest, given that it was the fastest to be executed and the majority of the students score the maximum points. Also, it may indicate a concept well understood by the students. When confronting the

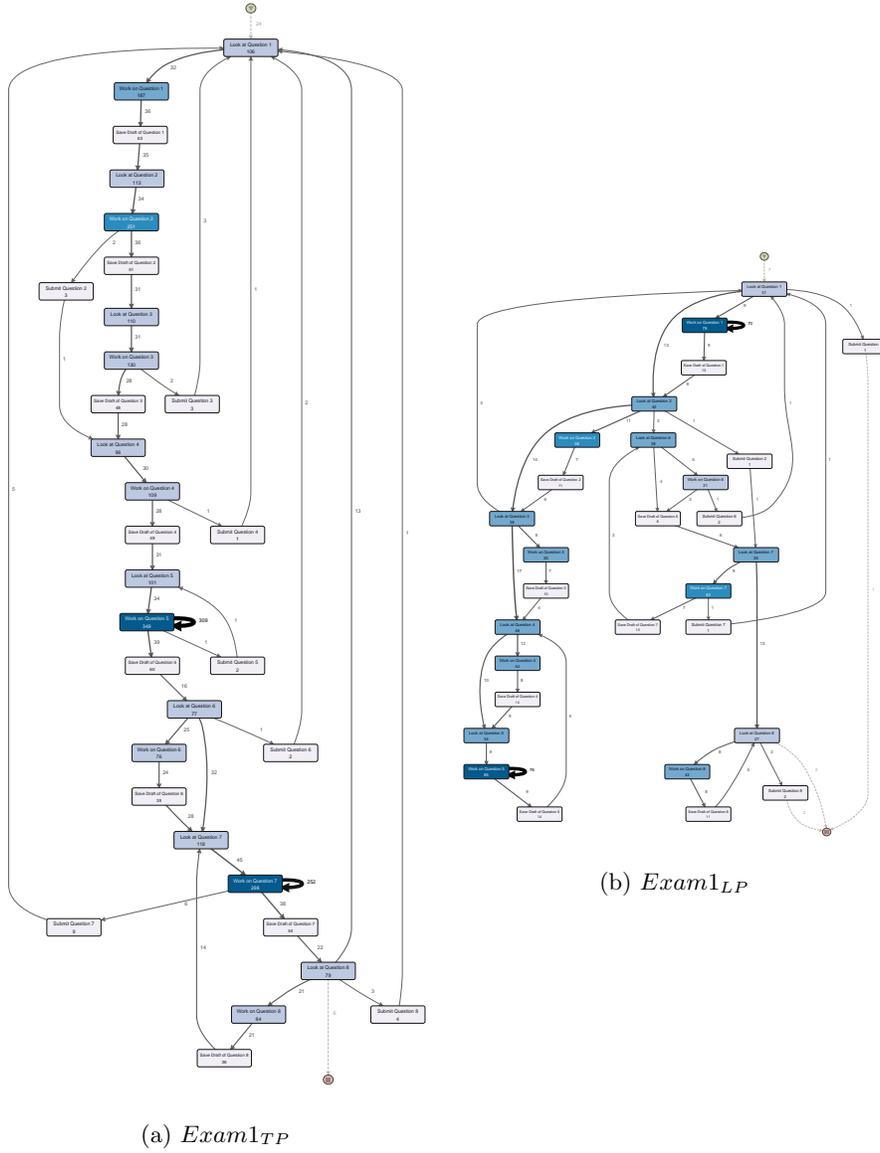


Fig. 2: Exam solving processes for the *Exam1* exam. Left (a): top performers (students who achieved more than 80 points). Right (b): lowermost performers (students who achieved less than 50 points)

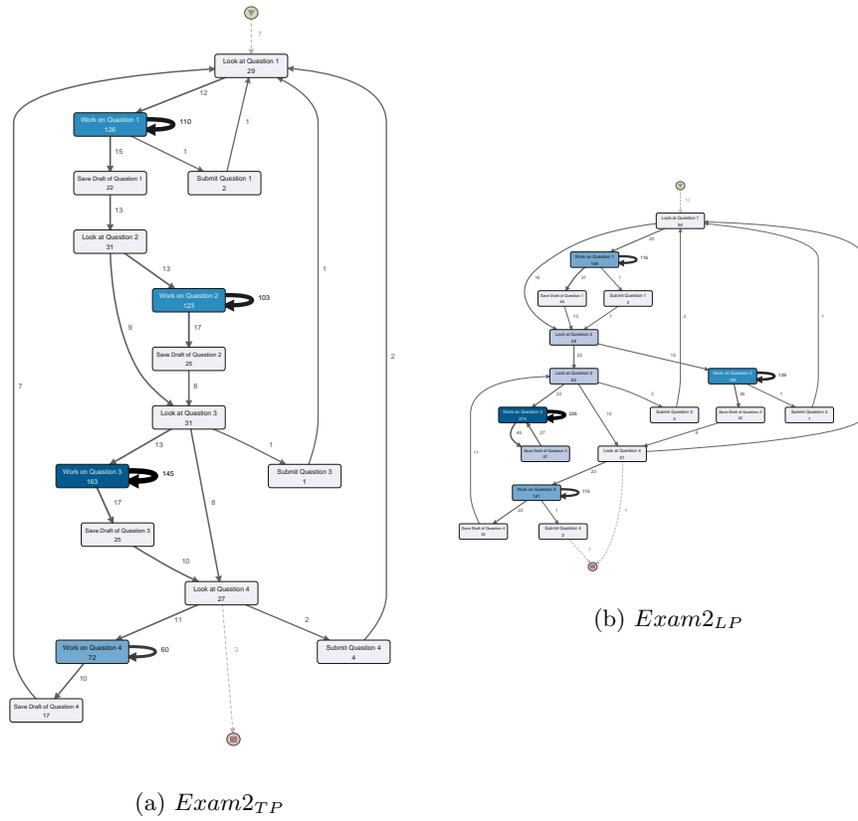


Fig. 3: Exam solving processes for *Exam2*. Left (a): top performers (students who achieved more than 80 points). Right (b): lowermost performers (students who achieved less than 50 points)

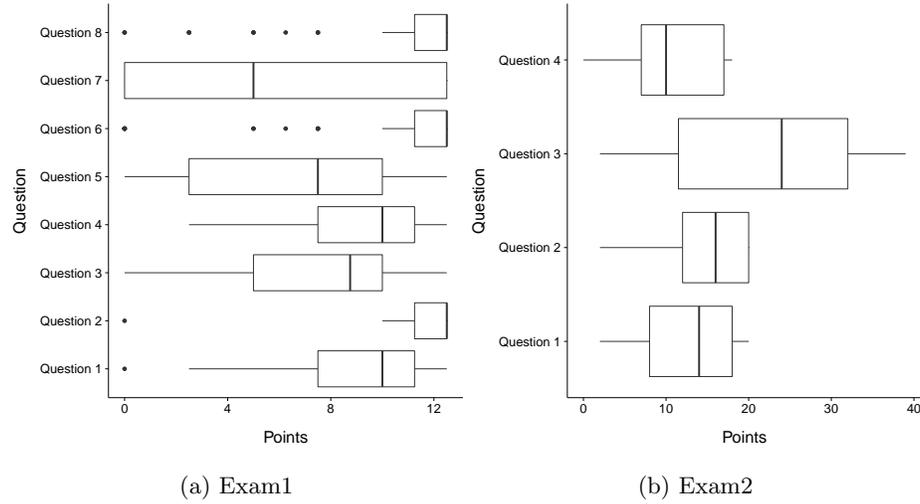


Fig. 4: Questions versus points achieved on them

data for Question 5, it was the second question with more spread points and given that it was the question with higher duration, it may indicate that it was a hard question or the concepts discussed in it were not well understood by the students. In *Exam2*, Question 2 was the question that demanded less time from the students. It was also the question where the students achieved highest points and there was a less variation on the points achieved. This result indicate to the instructor that either the question was easy or the concept discussed in it was well understood by the students. Question 3 had a duration higher than the other, but this was an expected result given that it was the biggest question. As the maximum points of this question is double the points of each of the other questions, it was also expected to be normal a duration of double the duration of each of the other questions. However, considering the medians this situation was not observed. Given the variation on points achieved it seems that the content of this question was not fully understood by some of the students.

A correlation analysis between the duration and the points for each question using Pearson correlation showed that only Question 2 in *Exam2* presents a correlation between these two attributes considering 95% confidence level. Correlation 0.418 and p-value 0.0299. In *Exam1* questions 1,2,3,5 and 6 showed a correlation between the two attributes. The correlations and p-values were (0.461, 0.000185), (0.289, 0.0241), (0.399, 0.00144), (0.492, 0.0000572) and (0.347, 0.00617) respectively.

4.3 Discussion

From the results observed it is not possible to state that there is not a common behavior on solving the exams (*Q1*). Each student created their own strategy for

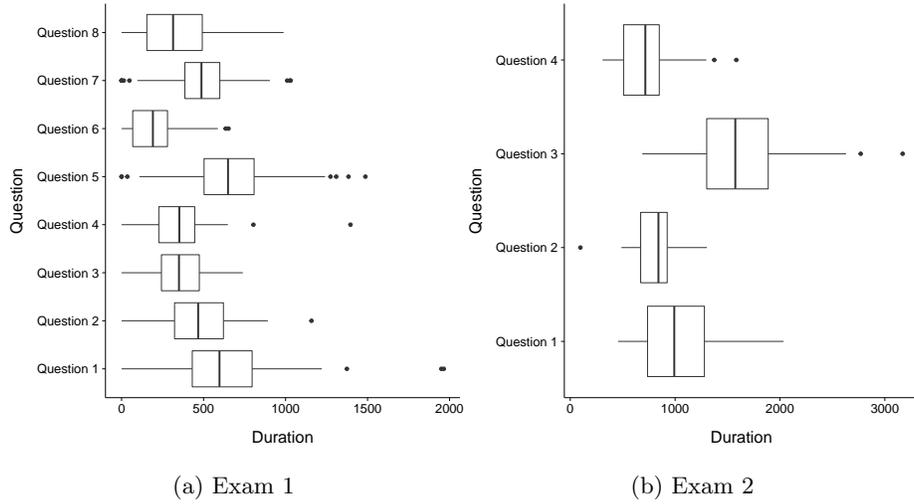


Fig. 5: Questions versus time (in seconds) spent on solving them

solving the exam. However, when evaluating groups of students based on their performance, it was possible to observe a pattern in the behavior of the topmost performers students. This group solved the questions in the same order in which they were presented in the exam.

When evaluating the correlation between the time spent for answering the questions and the points received for each of the questions only Question 2 in *Exam2* showed a correlation. This result conforms with the previous analysis that showed that the highest points were achieved in Question 2 and this question was the one with lower duration. It means that the students were confident when answering this question. It seems that its content was well assimilated by the students.

5 Conclusion

Process mining has been used in the education area to support the analysis of the behavior of students in online educational environments. In this paper, we used process mining to analyze the behavior of students when solving an exam with open ended questions. Trace data generated from the online teaching and learning environment was used to generate a event log. This event log was enriched with attributes that encoded the points received in each question and the time spent by the students to solve the questions.

We applied our approach in two exams performed by students from two master courses of an Austrian university. The results raised interesting questions for the instructor to investigate further, which may support them on the design of future editions of the courses. Especially, when designing further editions of the course, our method can help at better content and granularity of the questions.

Future work shall increment the depth and scope of the analysis of the educational data at hand. More specifically, we want to improve our analysis in two ways. First, we want to improve the encoding and consider more cases for the analysis, such as for example differentiating between students that receive zero points in a question because they answered it wrongly from those that did not answer the question. Second, we want to apply other kind of process mining techniques, such as conformance checking, in order to quantify how much deviation is associated to a good or a bad grade.

References

1. N. Trcka, M. Pechenizkiy, and W. van der Aalst, "Process mining from educational data," *Handbook of educational data mining*, pp. 123–142, 2010.
2. C. Domínguez, F. J. G. Izquierdo, A. J. Elizondo, B. Pérez, Á. L. Rubio, and M. A. Zapata, "Using process mining to analyze time distribution of self-assessment and formative assessment exercises on an online learning tool," *IEEE Trans. Learn. Technol.*, vol. 14, no. 5, pp. 709–722, 2021.
3. A. Bogarín, C. Romero, R. Cerezo, and M. Sánchez-Santillán, "Clustering for improving educational process mining," in *LAK*. ACM, 2014, pp. 11–15.
4. V. Aisa, A. P. Kurniati, and A. W. Yanuar Firdaus, "Evaluation of the online assessment test using process mining (case study: Intensive english center)," in *ICoICT*, 2015, pp. 472–477.
5. L. Ali and N. A. Dmour, "The shift to online assessment due to covid-19: An empirical study of university students, behaviour and performance, in the region of uae," *Int. J. Educ. Technol. High. Educ.*, vol. 11, pp. 220–228, 2021.
6. N. Berente, S. Seidel, and H. Safadi, "Research commentary - data-driven computationally intensive theory development," *Inf. Syst. Res.*, vol. 30, no. 1, pp. 50–64, 2019.
7. W. M. P. van der Aalst, *Process Mining - Data Science in Action, Second Edition*. Springer, 2016.
8. A. Bogarín, R. Cerezo, and C. Romero, "A survey on educational process mining," *WIREs Data Mining Knowl. Discov.*, vol. 8, no. 1, 2018.
9. P. Mukala, J. C. A. M. Buijs, M. Leemans, and W. M. P. van der Aalst, "Learning analytics on coursera event data: A process mining approach," in *SIMPDA*, ser. CEUR Workshop Proceedings, vol. 1527. CEUR-WS.org, 2015, pp. 18–32.
10. A. Baykasoglu, B. K. Özbel, N. Dudakli, K. Subulan, and M. E. Senol, "Process mining based approach to performance evaluation in computer-aided examinations," *Comput. Appl. Eng. Educ.*, vol. 26, no. 5, pp. 1841–1861, 2018.
11. M. Pechenizkiy, N. Trcka, E. Vasilyeva, W. Van der Aalst, and P. De Bra, "Process mining online assessment data." ERIC, 2009.
12. M. Bannert, P. Reimann, and C. Sonnenberg, "Process mining techniques for analysing patterns and strategies in students' self-regulated learning," *Metacognition and learning*, vol. 9, no. 2, pp. 161–185, 2014.
13. J. Saint, A. Whitelock-Wainwright, D. Gasevic, and A. Pardo, "Trace-srl: A framework for analysis of microlevel processes of self-regulated learning from trace data," *IEEE Trans. Learn. Technol.*, vol. 13, no. 4, pp. 861–877, 2020.
14. H. M. W. Verbeek, J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, "Xes, xesame, and prom 6," in *CAiSE Forum 2010*, ser. LNCS, P. Soffer and E. Proper, Eds., vol. 72. Springer, 2010, pp. 60–75.