CLUSTERING TECHNIQUES TO INVESTIGATE ENGAGEMENT AND PERFORMANCE IN ONLINE MATHEMATICS COURSES

Francesco Floris[1], Marina Marchisio[1], Fabio Roman[1], Matteo Sacchet[1] and Sergio Rabellino[2]

[1]Department of Molecular Biotechnology and Health Sciences, Via Nizza 52 - 10126 Turin, Italy

[2]Department of Computer Science, Corso Svizzera 185 - 10149 Turin, Italy

ABSTRACT

Among the various kinds of learning analytics emerging especially in the latest decade, clicking patterns cover a prominent role, fostered by their success in analyzing several types of data concerning activity on the web. They can be defined as sets of clicks performed by users, in which every set is treated as the basic unit. Few research has been performed on clicking patterns in educational contexts. In this paper, we perform analysis regarding clicks to an online course in Mathematics, aimed at allowing students to follow courses at a distance, both before and after enrolling at University. We used clustering techniques on students learning behavior, which have been defined for this research as visualizations of activities and resources of the course, to detect differences on students' grade according to their online learning behavior. Our results show that students tend to proceed on the course in both activities and resources. There is no correlation between participation and course grades, even if the most active students show higher scores. Moreover, patterns differ significantly according to the degree program of each student, showing the importance of tailored path.

KEYWORDS

Clicking Patterns, Clustering, Digital Education, Learning Analytics, Mathematics Teaching, Open Educational Resources, Open Online Courses

1. INTRODUCTION

The role of learning analytics (Siemens, 2012) has been widely established in studying students' behavior inside online courses. Since big data are extensively used as tools to perform analyses in a plethora of contexts, it is natural to consider their use in educational settings. Indeed, they allow optimizing courses under various aspects, such as structurally and strategically (Marchisio et al., 2019a; Barana et al., 2019). The importance of similar approaches has further increased in the latest few years due to the changes the COVID-19 pandemic required, when modalities such as blended (Ossiannilsson, 2017; Marchisio et al., 2020) and hybrid (Raes et al., 2020; Marchisio et al., 2022) became of common use, with the concrete perspective to retain them also after the definitive end of the emergency.

An analytic that is gaining interest over the years, also because it has already been experimented in other fields, are the click patterns. We can define them as the results of capturing the relationship between clicks, by treating the set of those a user performs as a single unit. Before their consideration in education, they have been widely used for instance as part of web analytics, since they allow a rich representation of mixtures of multiple navigational and informational intents (Duan et al., 2012). The main concept motivating their use is that each click, being it to a web result in a search engine output, or to educational resources and activities in the context of didactics, does not represent in general one unique intent. On the contrary, the intent of the user is often to explore many elements, so it is not possible to assume the presence of a one-to-one relation between clicks and intentions. Therefore, it could be important to aggregate actions to detect the actual presence of these complex mixtures of intents, and to give them, after being identified, a characterization with a meaning in the field of interest, in our case education.

In this paper, we apply click patterns to an online course in Mathematics simply called *Matematica in e-Learning*. This course falls within the *Start@UniTo* project (Marchisio et al., 2019b), started in 2017 and allowing students to follow online courses even before entering University, while they are still high school students. This enables brilliant students to take one or more exams in advance with respect to the first session traditionally available to first year students, thus starting their career in a comfortable way. The courses, which are opened also to university students that wish to study independently, cover disciplines from basically all the usual areas of knowledge: sciences (exact, life, social), humanities, languages, law studies, and so on. Some of them are taught entirely in English, for the sake of internationalization.

The course of *Matematica in e-Learning* is composed of twelve modules, each one corresponding to a topic of Mathematics generally taught during the first years of university in most scientific tracks: they include sets and numbers (01), linear algebra (02), univariate calculus (03-05), differential equations (06), bivariate calculus (07), probability and statistics (08-10), vector-valued functions (11), and numerical analysis (12).

It is usually addressed to students following a path in sciences, but not in the *hardest* STEM disciplines, since a student in Mathematics, Physics or Computer Science requires a more specific mathematical knowledge than what this course can grant. On the other hand, this course can be of interest for students following nonscientific paths since the importance of Mathematics extends also in everyday life and in forming conscious citizens. In this regard, it must be noted that the course is modular: it is not mandatory to require the students to study all the modules, but rather just some of them can be selected for specific paths. Indeed, several paths agreed with the university to consider *Matematica in e-Learning* as a course which is part of their curricula, often as an alternative to the course in Mathematics already present.

This research aims at investigating the differences in students' performance compared to their path in viewing activities and resources. Activities are interactive contents, those who require students to insert some inputs and receive outputs, usually a grade. Resources on the other side are transmissive contents (videos, books, pages, documents). The online course *Matematica in e-Learning* contains 58 activities and 193 resources. Students do not have to view and complete all these contents, the total number for each student depend on their program. The paper is structured as follows: Section 2 deals with some theoretical framework, while Section 3 presents the research questions and describes the methodology used. Section 4 reports all the results and discusses them; finally, some concluding remarks constitute Section 5.

2. THEORETICAL FRAMEWORK

In the last decade, several studies considered click patterns as a tool to analyze and predict students' behavior in the framework of an online course. (Sinha et al., 2014) studied clicks on videos, by considering interactions such as pressing “play” and “pause” or seeking and scrolling forward and backward. They were used to try devising predictions about how long students' interactions within the video lectures were, how much they felt engaged, and which patterns could be profiled as leading to in-video dropouts or stopping to view subsequent videos. (Crossley et al., 2016) combined such an approach with a natural language processing (NLP) analysis; this highlights the feasibility of studying clicks through different disciplines, and in conjunction with other automatic tools giving more points of view. (Tellakat et al., 2019) correlated clicks with psychological and psychometric aspects, but also suggested a wider perspective, by considering peculiarities like stability of clicks during time, temporal significance relative to when they click and study, how the way they click and study influences their final grade. (Rizvi et al., 2020) considered also how students relate with features allowing them to track their progresses via manual intervention designed to require awareness of the learning path. They were divided into three groups, based on how much they marked their activities as completed (all of them, only a part, none of them), and clicks were studied by taking account of this subdivision.

These studies fit into the challenges regarding the use of artificial intelligence (AI) in education (UNESCO, 2019). Two of them are of prominent interest here: the development of quality and inclusive data systems and, in turn, to make research on AI in education significant. Indeed, to possess reliable and timely data is necessary to have algorithms generate correct outputs, since predictions can be complete and accurate only if also the data on which they based are. Therefore, it is important to devote efforts in improving the techniques of analysis themselves too, because an attentive automatization implies the ability to retrieve more and better data in less time. This would allow students to learn better, more, and differently: personalization can be enhanced, better outcomes can result from their learning, and some learning goals unattainable without the use of technology can be obtained, an aspect which some models such as the SAMR (Hamilton et al., 2016) already highlighted.

As for other groundbreaking fields, it is still difficult to devise a theory in the strict sense of the term, since various research groups are developing the matter in different directions, thus requiring yet more time to identify that organic unity which allows for a precise formalization of all the aspects. This does not exclude these researchers from blazing a trail on which to base the search for evidence. In the case of our study, the relations between clicks, time, and grades, are of inspiration, as for some of the authors previously cited, but jointly with methods that we experimented in the context of other research projects. For example, we consider how the digital learning environment we use allows for the insertion of interactive worksheets designed with the Advanced Computing Environment (ACE) *Maple*, which fosters more interactivity and problem-solving skills in students (Fissore et al., 2021). This suggests considering how much students considered worksheets with respect to the overall resources and activities explored, as one of the characteristics useful to draw profiles of the students themselves.

3. RESEARCH QUESTION AND METHODOLOGY

Our research has the goal to find answers to the following questions:

- (RQ1) Which trend exhibits the clicking activity of students? Is it substantially uniform, or on the contrary there is some significance in identifying lower or higher online activity?
- (RQ2) If students use different patterns to learn, are there implications on their performances?
- (RQ3) Which specific patterns characterize the learning behavior of students from different degree programs?

This investigation is divided in two parts. At a first glance, we collected macro data on the students' attendance to the course, to provide an overview of the way students navigate the various type of contents. We considered a subset of students for answering, since we need their final grades for measuring performance, and a certain quota of course accesses and completion to look for statistically significant relations. These students are those who obtained a grade higher than 24/30. We compared on the one hand the final grades, and on the other hand the ratios between how many times students' opened *Maple* worksheets inside the course and the course completion percentages. The idea to construct such an indicator was fostered by observing that these worksheets, being based on an ACE, are more interactive than most of the other materials in the course, thus suggesting looking for a correlation between grades and proportion of study on resources with which students are well able to interact. In the second part, we considered the activity of students who completed the course, which means those who have a grade on the final test, a much smaller number of students compared to the total amount of subscribers, due to the open nature of the course. We decide to limit our analysis on this subset since students who completed the course should have a defined and complete learning pattern, which we defined by two numbers. The first is the percentage between the resources viewed by the user over the total number of resources that each student has to study. The second is the percentage between the activities viewed by the user over the total number of activities that each student has to study. Activities, in this setting, means interactive contents that require students to explore or insert some input. We then performed a clustering on the sample of students using the aggregative k-means algorithm. For the analysis, we used KNIME Analytics Platform 4.6.0. Data were collected from different databases and different systems, since the Digital Learning Environment that hosts the online course makes use of integrated tools. Data were cleaned, even for excluding resources and activities that should not be considered, such as information pages, evaluation questionnaires, activities for the issue of the certificate. Moreover, from the point of view of grades, the highest scores were considered, since students can attempt tests more than one time, even just for studying and reviewing.

4. RESULTS AND DISCUSSION

The overall sample of students who self-subscribed to the course is huge: 4,869 students entered the course at least once, with a slight majority of students from University of Turin (2,665). Globally, as can be seen in Figure 1, there is a similar trend regarding readings (darker line) and interventions (lighter line), with a different order of magnitude. The peaks mainly appear during winter exam sessions (mainly, from December to February), but not during the summer exam sessions. Furthermore, the longest period in which values remained high was the winter part of the academic year 2020-21, when a long wave of the COVID-19 pandemic forced institutions to provide minimal didactics in presence, and so online courses had one more reason to be useful. As a matter of fact, there is no uniformity, even just substantial: the periods of higher activity can be identified in a significant way.

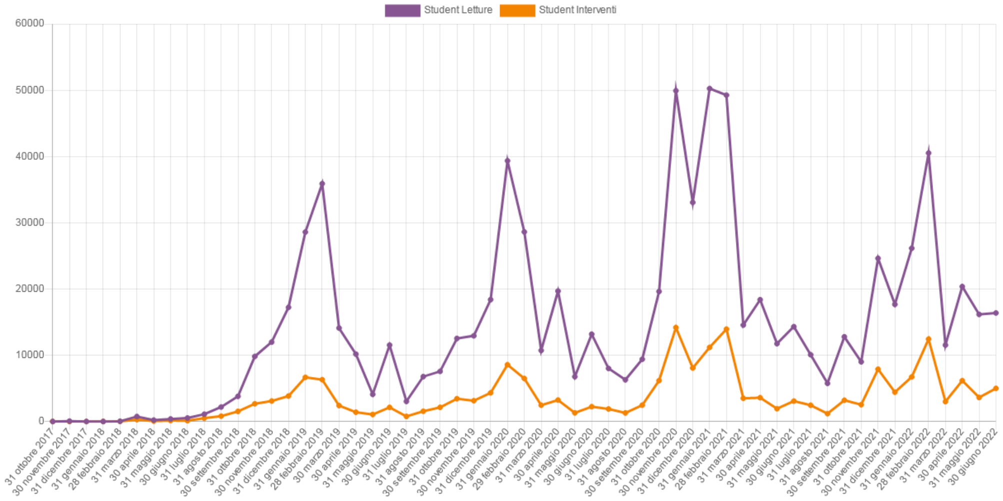

Figure 1. Click activity: Readings and interventions during time

Moreover, Figure 2 shows the scatterplot of ratios of openings of Maple worksheets versus grades. We see that the regression line, depicted as dashed in the plot, does not fit very well with the data: indeed, the Pearson correlation coefficient (PCC) is just about 0.09, highlighting a weak positive relation between the two variables. Another remark regards the quadrants in which the scatterplot is divided: the vertical and the horizontal line performing the division are respectively the median of the ratio and the median of the grades. Under the hypothesis of independence between the two variables, every quadrant should contain 25% of the data. The lower-right quadrant, referring to students obtaining grades below the median despite having a ratio above the median, contains in fact a 20% of the data, a lower value. Albeit not strong, this can be seen as evidence of the fact that accessing the ACE worksheets to a greater relative extent with respect to all the resources and the activities accessed helps to achieve a better performance.

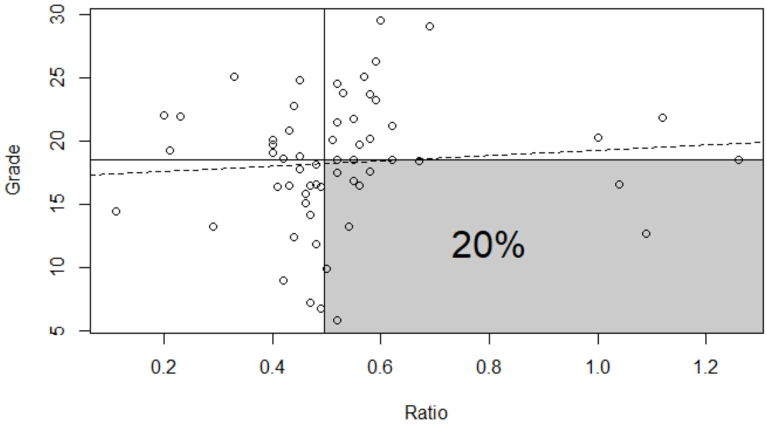

Figure 2. Ratios between ACE worksheet openings and completion percentages, versus final grades

Another statistic considers ordering grades decreasingly and counting how many ratios lie above the median for a certain number of the top grades: under the hypothesis of independence, they should be the half, so for example five in the top 10 and ten in the top 20. Table 1 shows the situation in our case: as depicted, for all the top 10, 20, and 30, the actual ratios above the median are more than half, thus establishing a high-high trend (which in turn implies a low-low trend), rather than exhibiting an independent behavior. Analogously, we can order ratios decreasingly, and count how many grades lie above the median for a certain number of the top ratios: again, Table 1 depicts how for all the top 10, 20, and 30, the actual grades above the median are more than half, so the previous considerations still hold.

Table 1. Number of ratios above median among best grades and number of grades above median among best ratios

Best grades	Expected ratios above median	Actual ratios above median	Best ratios	Expected grades above median	Actual grades above median
Top 10	5	8	Top 10	5	6
Top 20	10	14	Top 20	10	14
Top 30	15	18	Top 30	15	18

Table 1 confirm the presence of a positive relation between preferring ACE worksheets and obtaining better grades at the exam, which is not powerful, but it is also not negligible. Different learning patterns imply changes in how students perform, even if not extremely marked.

For the second part of the analysis, we collected the data of 237 students who completed the course (they submitted the final test). Figure 3 shows the scatterplot of users according to usage of resources and activities (scaled with respect to the different program that each student must follow) and the clustering output (the color of the points). Indeed, to grant a substantial equivalence between the *Matematica in e-learning* course and its in-person equivalent course (and thus avoiding students to select the simpler one, rather than basing the selection on the preferred modalities), the topics to be declared part of the exam had to be tuned on each path. For example, in the paths where probability and statistics were already part of the program, the relative modules were present at the exam, while in the paths lacking the (compulsory) study of probability and statistics, these modules were not included in the exam. Of course, students have in all cases access to every module, allowing them to explore and study at their will, regardless of what is mandatory (and part of the exam) for their track.

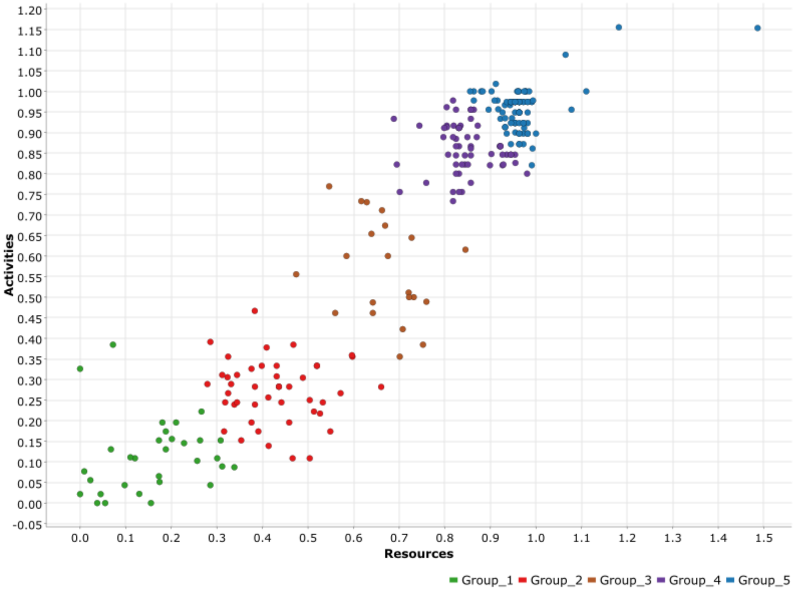

Figure 3. Scatterplot of students with respect to resources and activities viewed divided into 5 groups

Most of the students lies on the diagonal, meaning that they tend to perform in the same way activities and resources, the PCC is 0.8, a strong correlation. It must be noted that, even if the values of the grids are in percentages, there are some points which lie above 1 in both directions: this is due to the fact that the online course is completely open, so even if students choose the degree program in the online course and the system shows them just the related contents, they can anytime modify their choice and attend more than expected. The analysis considered just the last degree program choice of the user. Clustering algorithm divided the sample of students into 5 groups which differ in general for the number of online contents they viewed. After checking the goodness of the algorithm for different number of partitions, we opted for 5 clusters since it is a good turning point for the Sum of Squared Error (SSE), with more than 5 partitions the reduction of the SSE is not significant.

Moreover, as it can be seen from the boxplot in Figure 4, going from the group with less active students to the group with most active students there is a general tendence of increasing scores, in the boxplot the displayed scores are the median score and the quartiles. Students in Group 1 have larger dispersion since excellent students who do not need too much study belong together with less active and performing ones. Dispersion has a general tendence of decreasing, too. Every group contains low and high performing students, the difference is in the numerosity.

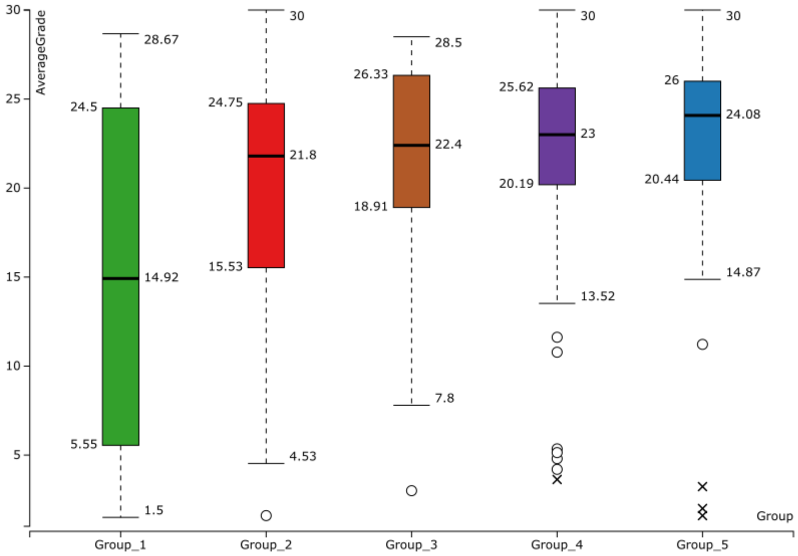

Figure 4. Conditional boxplot showing the distribution of grades with respect to the 5 clusters

Table 2. Number of ratios above median among best grades and number of grades above median among best ratios

Degree programs	Group 1		Group 2		Group 3		Group 4		Group 5	
	Count	Average	Count	Average	Count	Average	Count	Average	Count	Average
Biotechnology	0	0.0	2	26.4	1	24.4	2	24.6	0	0.0
Chemistry and Chemical Technologies	18	17.8	18	16.7	2	15.1	3	21.7	8	17.8
Pharmaceutical Chemistry and Technology	1	3.8	4	16.1	0	0.0	9	18.0	3	24.3
Pharmacy	1	27.8	0	0.0	5	23.1	2	29.0	2	24.4
Materials Science and Technology	2	3.7	0	0.0	0	0.0	0	0.0	1	20.6
Biological sciences	0	0.0	0	0.0	0	0.0	1	21.0	3	21.4
Agricultural Sciences and Technology	1	28.0	3	23.8	4	20.6	6	25.2	24	21.6
Forestry and Environmental Sciences	2	4.8	0	0.0	0	0.0	0	0.0	8	23.6
Natural Sciences	5	12.8	15	21.2	9	21.6	34	21.3	16	25.1
Strategic Sciences	0	0.0	0	0.0	0	0.0	0	0.0	1	25.2
Food Technology	1	6.8	2	24.3	0	0.0	1	19.1	11	24.3
Viticulture and Enology	0	0.0	1	26.0	0	0.0	1	28.0	4	24.2
Total / Average	**31**	**15.08**	**45**	**19.62**	**21**	**21.27**	**59**	**21.67**	**81**	**22.80**

Since students from different degree programs can attend the online course, we checked the number of students in the clusters for each degree program, as it can be seen in Table 2.

There is a large difference in the behavior of students according to their program. Students in Agricultural programs (Agricultural Sciences and Technology, Forestry and Environmental Sciences, Food Technology, Viticulture and Enology) who have the same study materials tend to view and complete most resources and activities since most of them are in Group 5. This is tendentially also true for students in Natural Sciences, where most of them are in Group 4, with Group 5 following as the second greatest one. Moreover, Group 5 in Natural Sciences has the greatest average grade among programs with more than three students. Students in the pharmaceutical sector concentrate on Groups 3, 4 and 5, thus they are quite active in online course attendance. On the other side, students in Chemistry and Chemical Technologies belong more to Groups 1 and 2. Some courses do not present enough students, they will be considered in the future when more data will be collected. It can also be noted that the online course *Matematica in e-learning* is not used in the same way by students of different program, and this suggests an implementation of online contents that are more suited to the training needs of students.

5. CONCLUSION

We can see that different students behaved differently. Of course, every student has his or her favorite ways to study and learn, and so we cannot exclude the personal component from these considerations. Moreover, they can explore different kinds of objects, either during a single session or in diverse sessions. The investigation of clicking patterns allowed us to determine some information about students' engagement and how they perform during the course, consequently resulting in certain grades at the exam. The answer to RQ1, relating students' patterns to the discriminant values that we choose of resources and activities, shows a quite uniform trend: students tend to complete both activities and resources in the same relative amount. Moreover, as it was seen from the first part of the analysis, the uniformity does not hold with respect to time: there is higher students' activity in the cold months, while winter exams are in program, thus in future elaborations even the time of access could be considered. We answered RQ2 by considering the average grade for each of the group that were created by the clustering algorithm, containing students with similar patterns. As a general remark, students with more visualization of resources and activities tend to have higher scores in the final test, thus it is clear that online attendance is important for the preparation of students, even though it was not supervised due to the open modality of teaching. Other specific indexes considered the interactivity of the resources and activities students explored. We detected how the more they use ACE worksheets, the more they are prone to perform well, by obtaining good grades. Finally, to answer RQ3 we analyzed the presence on each group of students from the different degree programs: there is a tendency even in this case, students from the same degree program seems to belong to the same clusters.

As future work, we would like to extend our inquiry to include a more systematic analysis of those patterns with automatized tools. The main challenge lies in dealing with massive databases, to be queried with detailed requests not so simple to connect, giving results that can require much work for their interpretation. It is also possible to consider a larger number of students (not just those who complete the final test) and study and predict their behavior, according to these indicators and to other ones that can be extracted from the data, with more than two parameters. It will help gaining further evidence regarding how students click on online platforms, their engagement and performance.

Another analysis we would like to conduct is related to the relationship between course content. The goal will be to understand how students move from one resource to another (also from a time perspective) and try to find common paths. In this way, on the one hand, we can find possible marginal content (which is perceived as disconnected from others within the course), and on the other hand, identify paths of content enjoyment that can then be proposed to other students who are in similar situations.

REFERENCES

Barana, A. et al., 2019. Learning analytics to improve formative assessment strategies. *Journal of E-Learning and Knowledge Society*, Vol. 15, No. 3, pp. 75-88.

Crossley, S. et al., 2016. Combining click-stream data with NLP tools to better understand MOOC completion. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge (LAK)*, Edinburgh, UK, pp. 6-14.

Duan, H. et al., 2012. Click patterns: An empirical representation of complex query intents. *Proceedings of the 21st International Conference on Information and Knowledge Management (CIKM)*, ACM, Maui, USA, pp. 1035-1044.

Fissore, C. et al. 2021. Development of Problem Solving Skills with Maple in Higher Education. In: Corless R.M., Gerhard J., Kotsireas I.S. (eds.) Maple in Mathematics Education and Research. MC 2020. Communications in Computer and Information Science, Vol. 1414. Springer, Cham.

Hamilton, E.R. et al., 2016. The Substitution Augmentation Modification Redefinition (SAMR) Model: a Critical Review and Suggestions for its USE. *TechTrends*, Vol. 60, pp. 433-441.

Marchisio, M. et al., 2019a. Boosting up data collection and analysis to learning analytics in open online contexts: An assessment methodology. *Journal of E-Learning and Knowledge Society*, Vol. 15, No. 3, pp. 49-59.

Marchisio, M. et al., 2019b. Start@unito: Open Online Courses for Improving Access and for Enhancing Success in Higher Education. *Proceedings of the 11th International Conference on Computer Supported Education (CSEDU)*, Heraklion, Greece, Vol. 1, pp. 639-646.

Marchisio, M. et al., 2020. Teaching Mathematics in Scientific Bachelor Degrees Using a Blended Approach. *Proceedings of IEEE 44th Annual Computer Software and Applications Conference (COMPSAC)*, IEEE, virtual, pp. 190-195.

Marchisio, M. et al., 2022. Valuable features of hybrid teaching in a higher education context. *Proceedings of European Distance and E-Learning Network Conference*, EDEN, Tallinn, Estonia, in press.

Ossiannilsson, E., 2017. Blended Learning State of the Nation. *International Council for Open and Distance Education (ICDE).*

Raes, A. et al, 2020. A systematic literature review on synchronous hybrid learning: gaps identified. *Learning Environments Research*, Vol. 23, pp. 269-290.

Rizvi, S. et al., 2020. Investigating variation in learning processes in a FutureLearn MOOC. *Journal of Computing in Higher Education*, Vol. 32, pp. 162-181.

Siemens, G., 2012. Learning analytics: envisioning a research discipline and a domain of practice. *2nd International Conference on Learning Analytics and Knowledge (ACM)*, Vancouver, Canada, pp. 4-8.

Sinha, T. et al., 2014. Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL, Doha, Qatar, pp. 3-14.

Tellakat, M. et al., 2019. How do online learners study? The psychometrics of students' clicking patterns in online courses. *PloS ONE*, Vol. 14, No. 3, pp. 1-17.

UNESCO, 2019. Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development. *Working Papers on Education Policy*, 07.