Data Analysis Techniques Applied to the MathE Database [⋆]

Beatriz Flamia Azevedo[1,2,5][0000−0002−8527−7409], Sofia Romanenko[1,4][0000−0002−5450−5712], Maria de Fatima Pacheco[1,3,5][0000−0001−7915−0391], Florbela P. Fernandes[1,5][0000−0001−9542−4460], and Ana I. Pereira[1,2,5][0000−0003−3803−2043]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, 5300-252 Bragança, Portugal
[2] Algoritmi Research Centre, University of Minho, Guimarães - 4800-058, Portugal
[3] Center for Research & Development in Mathematics and Applications CIDMA, University of Aveiro, Aveiro, Portugal
[4] University of Coimbra, Coimbra, Portugal
[5] Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Bragança - 5300-253, Portugal.
{beatrizflamia, sofia.romanenko, pacheco, fflor, apereira}@ipb.pt,

Abstract. MathE is an international online platform that aims to provide a resource for in-class support as well as an alternative instrument to teach and study mathematics. This work focuses on the investigations of the students' behavior when answering the training questions available in the platform. In order to draw conclusions about the value of the platform, the ways in which the students use it and what are the most wanted mathematical topics, thus deepening the knowledge about the difficulties faced by the users and finding how to make the platform more efficient, the data collected since the it was launched (3 years ago) is analyzed through the use of data mining and machine learning techniques. In a first moment, a general analysis was performed in order to identify the students' behavior as well as the topics that require reorganization; it was followed by a second iteration, according to the students' country of origin, in order to identify the existence of differences in the behavior of students from distinct countries. The results point out that the advanced level of the platform's questions is not adequate and that the questions should be reorganized in order to ensure a more consistent support for the students' learning process. Besides, with this analysis it was possible to identify the topics that require more attention through the addition of more questions. Furthermore, it was not possible to identify significant disparities in the students behavior in what concerns the students' country of origin.

[⋆] This work has been supported by FCT Fundação para a Ciência e Tecnologia within the R&D Units Project Scope UIDB/00319/2020, UIDB/05757/2020 and Erasmus Plus KA2 within the project 2021-1-PT01-KA220-HED-000023288. Beatriz Flamia Azevedo is supported by FCT Grant Reference SFRH/BD/07427/2021.

Keywords: e-learning platforms · mathematics · active learning · education technologies · data analysis

1 Introduction

All actors in the educational process are aware of the need to improve the quality of lectures and intensify research on innovations that contribute to better engage students and lower failure rates in the discipline. Due to its cumulative nature, courses that rely on a strong mathematical core present enormous challenges both to professors and students: in mathematics, students learn that through adequate reasoning and relying on proper assumptions, they can arrive at results that are fully trustable and applicable in a wide variety of scientific and real-life contexts. Guiding the students to the appropriate degree of attainment, comprehension, and autonomy has long been one of the significant challenges for professors, including at the higher-education level. Poor performance, especially in introductory courses, is a massive concern that college mathematics lecturers face [10, 11].

Although lecturing, in an exposition-centered approach, has been the traditional way of teaching, there is theoretical evidence of the need for students to be more active in constructing their understanding [2, 9, 14, 19].

Active learning methodologies, grounded in the constructivist theory of teaching and learning that holds that humans learn by actively using new information and experiences and that reality is shaped by the experiences of the learner, can be a meaningful contribution to boosting the students' engagement. Some of the main features of the constructivist teaching practice, such as the encouragement of the students' autonomy, initiative, and dialogue among their peers and with the professors, promote a sense of personal agency since the students have control of their learning and, to some extent, to their assessment [5, 7, 12, 18].

Students retain much more if they are challenged to reflect on and do more than just passively receive information. Active learning interventions can include approaches as diverse as workshops, group problem-solving and team quizzes, worksheets or tutorials completed throughout the class, use of personal response systems ("clickers") displaying a graph with the responses (there are many online applications for this purpose, such as https://www.polleverywhere.com), moments of individual thinking alternating with small group activities, all subject to immediate feed-back, are all powerful techniques for helping students work through and understand and solve a problem and are among the evidence-based best practice in active learning, and lead to greater learning [6, 17]. Cooperative learning is a component of active learning that is worth highlighting: it refers to work developed by the students, organized into teams, in order to produce an outcome of some sort: a laboratory or project report, the design of a product or a process or, within the context of learning mathematics, the solutions to a set of problems. The dynamics of cooperative learning should encourage face-to-face interaction, interdependence, individual accountability, appropriate use of interpersonal skills, and several moments of self-assessment of team performance and dynamics. Extensive research has shown that, when compared to traditional pedagogical models, cooperative learning – when it is implemented adequately – leads to greater learning and development of communication and teamwork skills, such as leadership, project management, and conflict resolution. Furthermore, the characteristics described before, go in the same direction as the 4th goal of the Sustainable Development Goals (SDGs), it is quality education, that intends to ensure inclusive and equitable quality education for all and promote lifelong learning [8].

The MathE online teaching platform is in line with the described scenario: it provides educational resources for students and lecturers, covering the traditional mathematics contents of higher-education courses. By registering on the platform, the users' have access to a wide variety of educational resources such as videos, solved exercises, podcasts, or pdf files, as well as questions that allow the students to undertake self-assessment tests and the professors to perform evaluation. MathE also provides the users with a forum where the students can share questions and challenges, teaching each other and, therefore, being active agents in the construction of new skills and knowledge. Professors and researchers also benefit from the existence of their forum in the platform where there is in-depth peer-to-peer interaction for the exchange of expertise and knowledge [13].

The platform aims to offer a dynamic and engaging tool to teach and learn mathematics, relying on interactive digital technologies that enable customized study. The goal of this research is to analyze the data collected on the MathE platform, over the 3 years, the platform has been online. So, the aims consist of investigating the topics available on the platform that need to be restructured in terms of questions level and also analyzing the students' performance according to the countries they belong to. This information will be combined with the conclusion of previous works [4, 3], in which [3] had investigated the profiles of different groups of students exclusively in the Linear Algebra topic, and [4] analyzed the optimum way to reorganize the resources available on the platform into different levels of difficulty. The information acquired in this work will complement the conclusions obtained in both papers. It will help the platform developers to trace the future path to provide intelligence for the MathE platform since it is expected that shortly the MathE will be able to make use of intelligent mechanisms, based on optimization algorithms and machine learning, to make autonomous decisions, tailored according to the needs of each user.

The rest of this paper is organized as follows. In Section 2, the collaborative educational platform MathE is briefly described. Section 3 presents the methodology adopted. Section 4, describes the data collected throughout the time that the platform is online, that are analyze in this paper. The results and discussion obtained based on the data analyze is presented in Section 5. Finally, the conclusion and consequently the direction of future works are presented in Section 6.

2 The MathE Platform

The development of Information and Communication Technologies (ICT) facilitates access to education and made the learning process more accessible, effective. Promoting an e-learning method requires different types of resources, in particular digital and technological resources. MathE is an e-learning platform focused on the mathematical contents of higher education courses. On the platform, any student or professor, has free access to a collection of questions, videos, and other pedagogical materials related to mathematics at higher education level. MathE was developed and implemented by a consortium of seven institutional partners from five European countries: Polytechnic Institute of Bragança (Portugal), the Limerick Institute of Technology (Ireland), the University of Genova, Pixel (Italy), Kaunas University of Technology (Lithuania), Technical University of Iasi (Romania) and EuroED (Romania). Each partner institution has built a solid community of professors in the corresponding countries that has been actively collaborating and responding to the project's challenges.

The MathE platform comprises three main sections: the *Student's Assessment* section is subdivided into two subsections: *Self Need Assessment* (SNA) and *Student Final Assessment* (SFA); the students can self-evaluate their knowledge using the subsection SNA whereas, on the other hand, under SFA, the professors can organize online tests about selected topics in this subsection. On the *MathE library*, the users can access a collection of videos and additional resources about the topics covered by the platform. Finally, the *Community of Practice* provides a free forum where users can create and share their experience, knowledge and information: the students are invited to discuss related issues and challenges in the Students' Community, and the professors can build a solid network of learning and teaching practices in the Lecturers' Community.

Moreover, MathE also offers a YouTube channel where all the videos of the platform are available. There are two types of videos (both available in the platform and in the MathE YouTube channel): the ones that were selected from the internet by the MathE experts (all linked with the MathE platform) and others exclusively produced by the MathE consortium according to the platform's needs (provided on the MathE platform and MathE YouTube Channel).

The MathE platform is currently being used by a significant number of users: there are enrolled 1171 students of 15 nationalities – Portuguese, Brazilian, Turkish, Tunisian, Greek, German, Kazakh, Italian, Russian, Lithuanian, Irish, Spanish, Slovenian, Dutch and Romanian. There are also 99 professors from 12 countries and 49 higher education institutions registered. It is important to emphasize that, besides the users signed up in the MathE portal, there are users from countries like India, Philippines and Egypt, taking into account the information obtained from the YouTube channel. Fig. 1 illustrates the MathE presence around the world, that is, the countries where the MathE has, at least, one person enrolled – either a professor or a student.

Currently the platform has 1841 questions, covering the fifteen most classical mathematical topics addressed in graduation courses. The questions available are divided into two levels of difficulty (basic and advanced) – this categorization

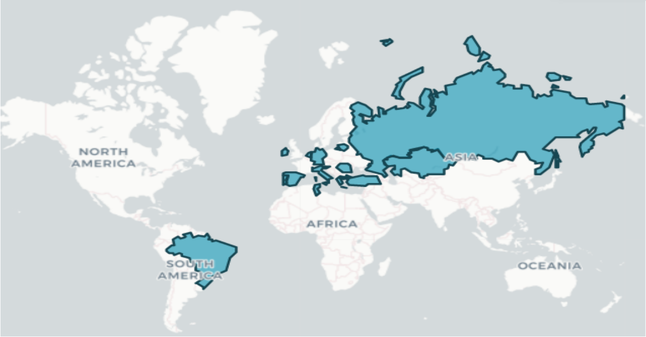


Fig. 1: MathE around the world

is done by a professor registered on the platform, both for the SNA and SFA sections. Table 1 describes the number of questions available in each topic at each section, SNA and SFA.

Table 1: MathE available questions according to topics and sections

Topics	**SNA Questions**	**SFA Questions**	**Total Questions**
1. Linear Algebra	211	101	312
2. Fund. Mathematics	327	43	370
3. Graph Theory	49	21	70
4. Differentiation	144	52	196
5. Integration	127	63	190
6. Analytic Geometry	40	20	60
7. Complex Numbers	41	20	61
8. Dif. Equations	41	20	61
9. Statistic	41	21	62
10. R. F. Single Var.	52	20	72
11. Probability	46	27	73
12. Optimization	96	37	133
13. R. F. Several Var.	58	22	80
14. Set Theory	40	19	59
15. Num. Methods	42	0	42
Total	**486**	**1355**	**1841**

It is essential to clarify that each time a student selects a topic and a question difficulty level to answer on SNA, a set of seven multiple-choice questions is randomly generated from an assessment platform database. After submitting the test for evaluation, the students will immediately receive feedback on their scores and some suggestions (extra material) will be given in the questions with the wrong answer. On the other hand, on the section SFA the quantity and which questions will compose the test are defined by the professor, who will schedule a test on the platform system composed by questions from an exclusive SFA database. In this case, after a student submits the test for evaluation, the professor immediately receives the student's score; the student only has access to their score 24 hours after the end of the test. Additional details about each section of the platform are described in [4, 3], and can also be found in its website (mathe.pixel-online.org) or at the MathE Youtube channel (MathE Channel).

3 Methodology

The methodology adopted in this paper consists in the application of strategies of data mining and machine learning to assess the data collected under the MathE platform. Using statistic tools, the data is analyzed with regard to student's hit probability for each performed topic. In this way, it is possible to identify the topic that requests more attention according to the student's difficulties to answer the available questions. Thereafter, the data is evaluated in compliance to the country of origin of the student, in order to search for different students' profiles according to their nationalities.

After that, the *k-means clustering algorithm* is used to identify the similarities and dissimilarities in the students' behavior, per country. Among the unsupervised methods, clustering techniques can be considered the most popular for grouping a set of elements with similarities in the same group and dissimilarities in other groups [15], an approach that is appropriate for exploring relationships between data and detecting the underlying structures.

The k-means partitioning clustering algorithm is one of the most well-known clustering algorithms. It consists of trying to separate samples into groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares (WSS) [1]. As k-means is not an automatic clustering algorithm, it requires the definition of the initial parameter k, that represents the number of clusters division. The value of k can be specified by different techniques, but in this work the *Silhouette method* [16], which is a similarity measurement, is adopted. Once this value is established, the k-means algorithm divides a set of X samples $X_1, X_2, ..., X_m$ into k disjoint clusters C_k, each described by the mean of the samples in the cluster, μ_i, also denoted as cluster "centroids". In this way, the k-means algorithm aims to choose centroids that minimize the inertia, or within-cluster sum-of-squares criterion, presented in Equation (1) [1].

$$WSS = \sum_{i=0}^{m} \min ||X_j - A_i||^2, \text{ in which } \mu_i \in C_k \tag{1}$$

From these centers, a clustering is defined, grouping data points according to the center to which each point is assigned. The k-means clustering algorithm and the Silhouette algorithm exist in the MatLab® library and they were applied in the research that this work describes.

4 Description of the MathE Data

MathE is an international platform and, currently there are 1171 students enrolled and 99 professors and researchers from 15 different nationalities. Table 2 describes this information with more detail; it is possible to observe that the countries with more users under the profile of student are Portugal, Lithuania and Italy.

Table 2: MathE users distribution according to their countries

Country	**Students**	**Professors**	**Institutions**
Portugal	646	33	18
Lithuania	231	21	11
Italy	171	13	3
Ireland	63	12	2
Romania	50	10	2
Slovenia	4	1	1
Tunisia	2	0	1
Spain	1	3	3
Netherlands	1	2	1
Turkey	1	1	1
Russia	1	1	1
Germany	0	0	1
Greece	0	0	1
Kazakhstan	0	1	1
Brazil	0	1	2
Total	**1171**	**99**	**49**

The data collected for analysis in this work considers information of 6927 answers distributed among the 15 topics of Table 1. These answers were provided by 284 students that uses the SNA section, since the platform' launch, in 2019. It is important to highlight that the questions and the topics are constantly being added to the platform, then, naturally some topics have more questions answered than others. Table 3 describes the MathE collected data.

In this table Fund. Mathematics, Dif. Equation, R. F. Single Var. and, R. F. Several, Var, means respectively Fundamentals of Mathematics, Differential Equation, Real Function of Single Variable, and Real Function of Several Variables. The data is fully characterized by the topic and the two levels of difficulty – basic and advanced. The *Topics* column describes all the MathE topics available on the platform. Moreover, in both levels (middle block and right block) the *Std* column shows the number of students that answered questions on that topic; the number of correct answers and incorrect answers is described in the columns *CA* and *IA*, respectively. Finally, the columns *TQ* present the total number of

Table 3: MathE collected data

Topics	**Basic Level**				**Advanced Level**			
	Std	**CA**	**IA**	**TQ**	**Std**	**CA**	**IA**	**TQ**
1. Linear Algebra	135	1353	1624	2977	56	363	428	791
2. Fund. Mathematic	45	332	364	696	3	33	32	65
3. Graph Teory	5	13	14	27	2	16	5	21
4. Differentiation	49	180	357	537	3	15	10	25
5. Integration	13	45	52	97	2	4	3	7
6. Analytic Geometry	23	139	143	282	6	17	33	50
7. Complex Numbers	29	78	173	251	22	130	105	235
8. Dif. Equations	11	50	42	92	0	0	0	0
9. Statistic	57	152	174	326	0	0	0	0
10. R. F. Single Var.	8	28	39	67	0	0	7	7
11. Probability	11	27	50	77	3	11	5	16
12. Optimization	3	4	21	25	0	0	0	0
13. R. F. Several Var.	3	5	13	18	0	0	0	0
14. Set Theory	3	22	13	35	0	0	0	0
15. Num. Methods	8	97	106	203	0	0	0	0
Total	-	**2525**	**3185**	**5710**	-	**589**	**628**	**1217**

questions answered at each level (by topic), which corresponds to the sum of all correct and incorrect answers of that difficulty level.

5 Data Analysis

In this section, the analysis of the data previously described is presented, which aims to investigate the students' behavior on the MathE Platform since it has been online. First of all, it is essential to clarify that, as mentioned above, there are 1171 students enrolled on the platform, of which 284 use the SNA section; the others students use other resources of the platform such as videos and/or pedagogical materials or, even, the community of Practice. These 284 students belong to 8 countries: Portugal, Lithuania, Italy, Ireland, Romania, Russia, Spain, and Slovenia. Considering the information these students provided, a global analysis of the data set is done after a complementary analysis by countries.

5.1 General Database Analysis

Initially, the global performance of the students on the platform was analyzed, that is, the data of students who used the platform to answer the questions available in the topics of the Student Need Assessment section (SNA). From Table 3, it is possible to see that the number of basic questions answered (5710 in total) represents 82% of the total questions answered on the platform against 1217 advanced questions answered. So, it is clear that the students prefer to utilize the basic questions more than the advanced ones. In terms of the type of answer, in general, the number of incorrect answers is higher than the correct ones, 3185 incorrect answers in the basic level (56%) and 628 $(51,6\%)$ in the advanced one.

Comparing the data presented in Table 1, which describes the number of questions available, and Table 3 that presents the number of questions answered by the students, it is possible to note that the topics most required by students are those with the great variety of questions available. It can be justified because the contents of the MathE platform are constantly updated, with additional questions on each topic. So, in Table 3, Linear Algebra is the most used topic, followed by Fundamentals of Mathematics, Differentiation, and Complex Number, with more than 450 answers in terms of total answers.

Therefore, to investigate the distribution of the hit obtained in each topic, the individual hit probability per topic is calculated for each student. The graphic results of this evaluation are presented in terms question level Basic (Fig. 2 – 275 students) in which it is intend to compare the probability of questions correctly answered on the 15 topics available on the MathE Platform.

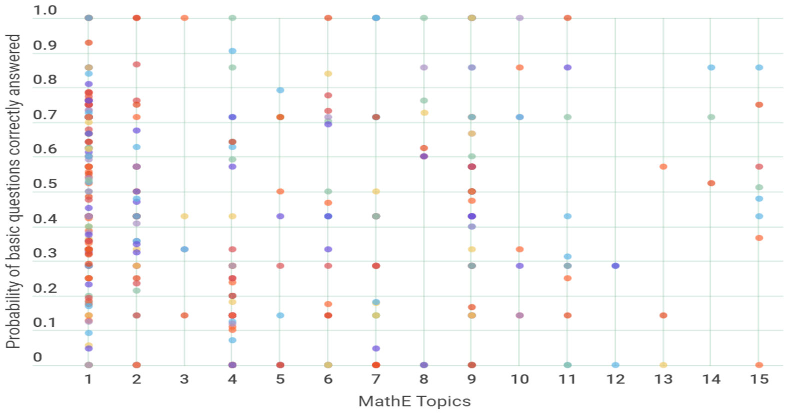


Fig. 2: Probability of correct answers per topic for all student who answered basic questions

As it is possible to see, in some topics the students distribution is almost homogeneous over the interval $[0, 1]$, which means there are both students with excellent performance (close to 1), and students with poor performance (close to 0), as well as students with average performance. These characteristics can be found in topics 1, 2, 4, and 9, which are the topics with more presence of students, which is denoted by the colorful points.

Nonetheless, in some topics, it is possible to observe the presence of gaps in the performance interval, which can also be associated with the low number of students that use the topic, but some peculiarities can be observed that lead us to meaningful observations. The topic 5 (Integration) is used by 13 students, only 2 have a performance greater than 0.5, whereas in the topic 7 (Complex Numbers), which 29 students use, only 1 student has a performance between the interval $]0.5, 1[$. The absence of students in these topics may indicate the lack of easier questions. On the other hand, in the topic 8 (Differential Equation), in which there are 10 students enrolled, there is no student with a performance in the interval $]0, 0.6[$, which indicates that is mandatory more complex questions in the basic level of this topic. Finally, on the topic 15 (Numerical Methods), the majority of students have a performance between $]0.35, 0.6[$, which indicates the necessity of questions with more variability in terms of difficulties.

Although all the questions belong to a basic level, some are more basic than others, while others have a more difficult degree, the questions are not all at the same level. The variability in difficulty within a given level is expected, and it is important to maintain this, considering that there are students with different needs enrolled on the platform. But some topics are not meeting this expectation, which calls for a better distribution of the questions in more levels of difficulty, as already indicated in [4]. Such observations are fundamental for the level of difficulty of the future questions that will be inserted in the topics; mainly, the topics 5 and 7 need easier questions, and the topic 8 requires more complex questions. In contrast, the topic 15 needs both of them.

Considering the few questions answered and also the few students practicing advanced questions, it is not possible to have consistent conclusions about the topics and the advanced level of the questions' difficulty.

5.2 Students Assessment per Country

As previously mentioned, in the SNA section, there are students from 7 countries, so in this section, the students' performance according to the countries is surveyed. Table 4 describes the data through countries in terms of the number of students per country that answered basic and advanced questions; and also in terms of the type of the answer (correct and incorrect) in both difficulty levels (basic and advanced). Finally, at the last column the sum of all questions answered is presented.

As can be seen, most students using the SNA are from Portugal, Lithuania, and Italy. These three countries have at least one institution on the platform's developer team, contributing to greater platform dissemination. Table 2 shows that the three countries have the most registered students, professors, and institutions on the platform. Besides, it is worth mentioning that Portuguese students correspond to practically half of the students enrolled in the platform 646 (out of 1171). Concerning the SNA section, Portuguese students are more than 60% of the total students, it is 174 out of 284.

Thus, to analyze the students' performance by country, the probability of correct answers for the questions by country was obtained and is shown in Table 5. Thus, the columns *Basic Questions* and *Advanced Questions* correspond to the hit average of the students in the basic and advanced levels, respectively.

Table 4: Students Performance (per country)

Country	**Number of Students**			**Basic Answers**		**Advanced Answers**		**Total**
	Basic	**Advanced**	**Total**	**Correct**	**Incorrect**	**Correct**	**Incorrect**	
Portugal	168	60	174	1529	1879	366	402	4176
Lithuania	53	11	55	486	634	113	148	1381
Italy	35	7	35	333	446	48	28	855
Ireland	12	3	13	100	138	30	19	287
Romania	3	1	3	18	14	17	11	60
Russia	1	1	1	30	54	11	17	112
Spain	1	0	1	12	16	0	0	28
Slovenia	2	1	2	17	4	4	3	28
Total	275	84	284	2525	3185	589	628	6927

Furthermore, the last column, *All Questions* is the hit probability considering both levels.

Table 5: Students hit average probability (per country)

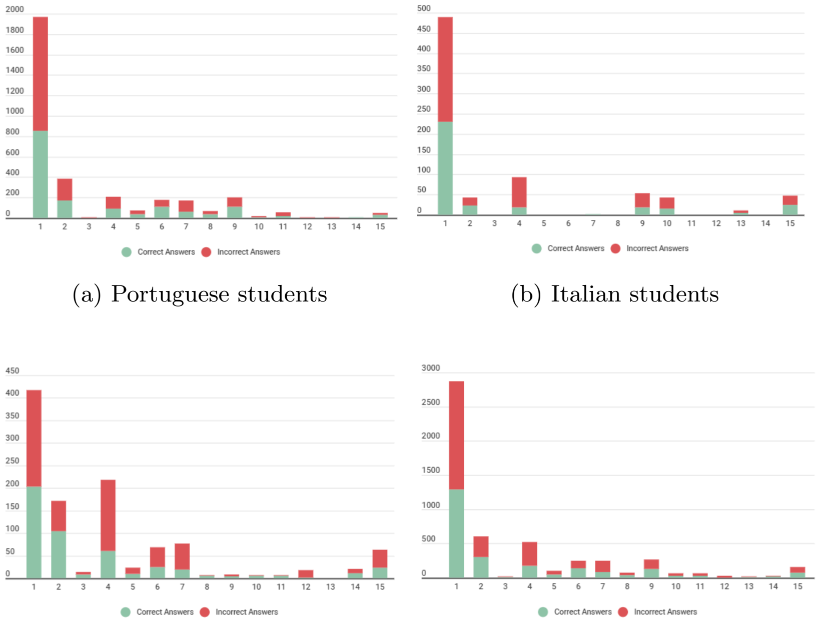
Country	**Basic Questions**	**Advanced Questions**	**All Questions**
Portugal	0.45	0.48	**0.45**
Lithuania	0.43	0.43	**0.43**
Italy	0.43	0.63	**0.45**
Ireland	0.42	0.61	**0.45**
Romania	0.56	0.61	**0.58**
Russia	0.36	0.39	**0.37**
Spain	0.43	0.00	**0.43**
Slovenia	0.81	0.57	**0.75**

From Table 5, it can be seen that the hit average for the advanced questions is almost always more significant than the probability of correct answers for the basic questions, and the opposite was expected, since in the basic questions, the students make many mistakes, so a low hit probability was expected at the advanced questions. This observation may indicate that the questions are not adequately organized on the platform since the basic questions have a degree of difficulty higher than expected and the advanced ones are not as complex as wished. Thus, for the best use of it, this issue is one of the urgent points to be reviewed for platform improvement.

The OECD Programme for International Student Assessment (PISA) examines what students know about mathematics, and according to this ranking, the countries and their mean classification are: Slovenia (509), Ireland (500), Portugal (492), Russia (488), Italy (487), Lithuania and Spain (481) and Romania (430) [11]. Since there is not an expressive number of students in all countries, it is not easy to establish a highly reliable comparison. Thus, in order to consider only countries with more than 30 students (Portugal, Lithuania, and Italy), it is noted that the average of both in PISA is close, with Portugal in 28th position, Italy in 31st and Lithuania in 34th [11]. Thus, it was already expected that the student's performance would be similar, as can be seen in the averages of correct answers in Table 5, mainly by the last column, which considers all the questions answered by the students of that country.

5.3 Portuguese, Lithuanian and Italian Students Assessment on Basic Questions

As already mentioned, there is a small number of students per country, so it is not feasible to assess the profile of students from the 8 countries. Therefore, this section will only consider data from Portuguese, Lithuanian, and Italian students, as there are more than 30 students in each group. Moreover, the number of advanced answers is few representative when compared to the basic answers. Therefore, only basic answers are considered in this section.



(a) Portuguese students

(b) Italian students

(c) Lithuanian students

(d) Students of the three countries

Fig. 3: Students' assessment on basic questions for each topic (per country)

As shown in Fig. 3, students from the three countries predominantly answer topic 1 - Linear Algebra. This one is widespread in practically all higher education courses, regardless of the country. This may be the main reason for such a significant number of answers.

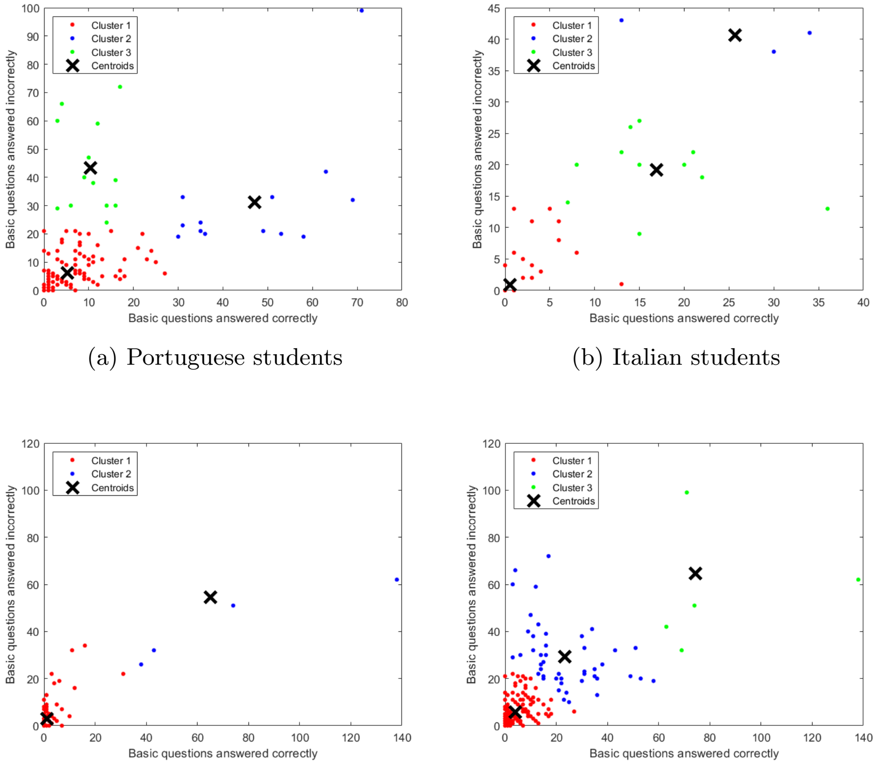
In respect to Portuguese students, in Fig. 3a, it is possible to verify that they pay extreme attention to topic 1 (Linear Algebra) reaching approximately 2000 responses, while all other topics have less than 210 responses (with the exception of topic 2 - Fundamentals of Mathematics). In addition to the already presented justification of Algebra Linear being present in several courses, the fact that the other topics are less used may be related to the encouragement given by the professors during the classes. Similar behavior is found in Italian students, Fig. 3b, in this case, while Algebra Linear collects almost 500 responses, the other topics have less than 100. On the other hand, in Lithuania, Fig. 3c this pattern is less expressive, and although with a smaller amount of answers than in other countries, the topics 2 – Fundamentals of Mathematics, 4 – Differentiation, 6 – Analytic Geometry, 7 – Complex Numbers and 15 – Numerical Methods, are also being significantly explored by the students in relation to the other topics used by the Lithuanian students.

The data collected presented on Fig. 3 is interesting and worthy to be explored in future works. If one can perceive strengths that lead students to have a preference for Linear Algebra, it will be possible to export this characteristics to others, thus captivating students to use the platform constantly and intensively in other topics too.

Finally, to identify the similarities and dissimilarities in the students' behavior, a clustering analysis was performed and the results are shown in Fig. 4.

Regarding the Portuguese students, in Fig. 4a, the algorithm grouped the students into 3 clusters. Thence, in cluster 1 (red), there are the students that answered fewer questions in relation to the other clusters, as it is the cluster with the highest population density. These students answer a maximum of 50 questions, which is represented by the sum of the x and y coordinates. Thus, the cluster 1 represents students who use less the platform, with mean equal to 12 answered and the students have an average performance in relation to the others. On the other hand, the cluster 2 (blue), are the students who answered more basic questions correctly. All students of this cluster answered at least 30 basic questions correctly and more than 18 incorrectly, while the majority answered less than 40 incorrectly. Furthermore, the average of answers is 78, so on cluster 2, students who use the platform more often and have to perform better than other students. Finally, in cluster 3 (green), there are students who also use the platform a lot, an average equal to 55 but do not perform well since they have a high error rate and a low rate of success in basic questions.

In the case of Italian students, Fig. 4b there are also 3 clusters, but with different behavior from Portuguese students. In the case of cluster 1 (red) of Italians, we have students who answer a few questions (mean equal to 2 and maximum of 17 question), and most of the answers are incorrect. In cluster 2 (blue), there are the students who answered the largest number of questions,



(a) Portuguese students

(b) Italian students

(c) Lithuanian students

(d) Students of the three countries

Fig. 4: Clustering of the basic questions answered (per country)

with mean of answers equal to 66, but the number of incorrect answers is much higher in relation to the number of correct answers. Finally, in cluster 3 (green), we have students with average performance, in this case the number of correct answers and errors is more balanced than the others, and these students answer an average of 36 questions.

For the Lithuania, Fig. 4c there are 2 clusters. Students are heavily concentrated in cluster 1 (red), responding to a few questions (means of answer equal to 5) with more incorrect than correct answers. Furthermore, in cluster 2 (blue), we have the students who answer the most questions (means equal to 133); however, this is a small group composed of 4 students, and although the performance is slightly higher than the students in cluster 1, it is still not excellent, considering the number of errors.

Finally, in Fig. 4d, there are the students from the 3 countries, cluster 1 are the students who answer fewer questions and with a low rate of correct answers; cluster 2, students who answer more questions than those in cluster 1 and less than those in cluster 3, but still with an intermediate performance. Moreover, in cluster 3, the students who answer more questions are represented by a small number of students.

6 Conclusions

The MathE platform is an online educational system that aims to help students who struggle to learn college mathematics, as well as students who want to deepen their knowledge of a multitude of mathematical topics, at their own pace. The platform currently provides a set a diversified questions, videos and pedagogical resources for the higher educational level. The question are randomly generated, independently of the profile of the users (there is only the possibility to choose topic, subtopic and level of difficulty), but it is expected that in the near future the platform will be able to make use of intelligent mechanisms, based on optimization algorithms and machine learning, to make autonomous decisions, able to direct the questions in a customized way, according to the students profile and needs.

The research [3, 4] aimed to investigate the difficulties and potentialities of the platform, as well as the characteristics that could be used to make the platform more efficient. Thus, the approach presented in this paper seeked to evaluate the adequate level of difficulty of the questions in the topics that are available on the platform, based on the students' hit probability at the SNA section. In addition, it was also evaluated whether the country of origin is a relevant variable in the students' performance. Thus, the information collected through this research will serve as a guide to make the choice of optimal strategies to improve the performance of the platform.

From the results obtained in this work, together with the others already carried out [3, 4], it is evident the need to reorganize the questions in more levels of difficulty. However, the results of this analysis will be fundamental for defining the type of questions that each topic needs. In addition, currently the assignment of a question to a certain level is done by a collaborating professor, so this division is subject to partiality and subjectivity, and may vary from person to person. Thus, finding a way to assign the questions to their respective difficulty level autonomously, through an intelligent system, is one of the possible ways to improve the organization of questions on the platform. This is also a way to keep students constantly active on the platform, as more engaged the students are in the platform uses, more questions they performs.

Finally, in relation to the analysis by countries, from the data analyzed so far, it is not possible to conclude whether students from a particular country perform better than from other countries (due, for example, to the quality of education in the country in question or other factors). In general, countries among the ones that have more than 30 students enrolled in the platform, show very similar outcomes in questions of both levels of difficulty. Thus, with the data that is currently have available, the country of origin does not appear as a determining variable in the customization of questions for students, in a future version of the platform.

References

1. Arthur, D., Vassilvitskii, S.: K-means++: The advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. p. 1027–1035. SODA '07, Society for Industrial and Applied Mathematics, USA (2007)
2. Ausubel, D.: The Acquisition and Retention of Knowledge: A Cognitive View. SPRINGER-SCIENCE+BUSINESS MEDIA, B.V (2000)
3. Azevedo, B.F., Rocha, A.M.A.C., Fernandes, F.P., Pacheco, M.F., Pereira, A.I.: Evaluating student behaviour on the mathe platform - clustering algorithms approaches. In: (In press) Book of 16th Learning and Intelligent Optimization Conference - LION 2022. pp. XX–XX. Milos - Greece (2022)
4. Azevedo, B.F., Amoura, Y., Rocha, A.M.A.C., Fernandes, F.P., Pacheco, M.F., Pereira, A.I.: Analyzing the mathe platform through clustering algorithms. In: Gervasi, O., Murgante, B., Misra, S., Rocha, A.M.A.C., Garau, C. (eds.) Computational Science and Its Applications – ICCSA 2022 Workshops. pp. 201–218. Springer International Publishing, Cham (2022)
5. Bransford, J., Brown, A.L., Cocking, R.R.: How people learn: Brain, mind, experience, and school. `http://www.nap.edu/books/0309070368/html/(accessed` (2000), accessed June, 2022
6. McKeachie, W.J.: Teaching Tips: Strategies, Research, and Theory for College and University Teachers. Houghton Mifflin: Boston, MA (2002)
7. MCLeod, S.A.: Constructivism as a theory for teaching and learning. `www.simplypsychology.org/constructivism.html` (2019), accessed June, 2022
8. Nation, U.: Sustainable development goals. www.un.org/sustainabledevelopment/education/ (2020), accessed August, 2022
9. Novak, J.D., Gowin, D.B.: Learning how to learn. Cambridge Univ. Press (1984)
10. OECD: The future of education and skills - education 2030. `https://www.oecd.org/education/2030/E2030%20Position%20Pape%20(05.04.2018).pdf` (2018), accessed June, 2022
11. OECD: PISA 2018 Results (Volume I). OECD publishing (2019)
12. Oliver, K.: Methods for developing constructivist learning on the web. Educational Technology **40**, 5–18 (01 2000)
13. Pacheco, M.F., Pereira, A.I., Fernandes, F.: Mathe - improve mathematical skills in higher education. In: Proceedings of the 2019 8th International Conference on Educational and Information Technology (ICET2019). pp. 173–176. United Kingdom (2019)
14. Piaget, J., Claparède, E.: The language and thought of the child. Routledge (1959)
15. Rokach, L., Maimon, O.: Clustering methods. In: Clustering methods Data mining and knowledge discovery handbook, Maimon, O. and Rokach, L. (eds). Springer (2005)
16. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987)
17. Trujillo, G., Tanner, K.: Considering the role of affect in learning: Monitoring students' self-efficacy, sense of belonging, and science identity. CBE life sciences education **13**, 6–15 (03 2014)
18. Vygotsky, L.S.: Mind in society: The development of higher psychological processes. Harvard University Press (1978)
19. Vygotsky, L.S.: Thought and Language. MIT Press (1986)