Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2022, 10, 22-42
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Analyzing Differences between Online Learner Groups during the COVID-19 Pandemic through K-Prototype Clustering

Guanggong Ge[1,2], Quanlong Guan[1,2,3*], Lusheng Wu[4*], Weiqi Luo[1,2,3], Xingyu Zhu[4]

[1]College of Information Science and Technology, Jinan University, Guangzhou, China
[2]College of Cyber Security, Jinan University, Guangzhou, China
[3]Guangdong Institute of Smart Education, Jinan University, Guangzhou, China
[4]Department of Science and Technology, Jinan University, Guangzhou, China
Email: guanggongge@stu2019.jnu.edu.cn, *gql@.jnu.edu.cn, *twls@jnu.edu.cn, luoweiqi@jnu.edu.cn, 16828611@qq.com

How to cite this paper: Ge, G.G., Guan, Q.L., Wu, L.S., Luo, W.Q. and Zhu, X.Y. (2022) Analyzing Differences between Online Learner Groups during the COVID-19 Pandemic through K-Prototype Clustering. *Journal of Data Analysis and Information Processing*, **10**, 22-42.
https://doi.org/10.4236/jdaip.2022.101002

Received: December 22, 2021
Accepted: January 27, 2022
Published: January 30, 2022

Abstract

Online learning is a very important means of study, and has been adopted in many countries worldwide. However, only recently are researchers able to collect and analyze massive online learning datasets due to the COVID-19 epidemic. In this article, we analyze the difference between online learner groups by using an unsupervised machine learning technique, *i.e.*, k-prototypes clustering. Specifically, we use a questionnaire designed by domain experts to collect various online learning data, and investigate students' online learning behavior and learning outcomes through analyzing the collected questionnaire data. Our analysis results suggest that students with better learning media generally have better online learning behavior and learning results than those with poor online learning media. In addition, both in economically developed or undeveloped regions, the number of students with better learning media is less than the number of students with poor learning media. Finally, the results presented here show that whether in an economically developed or an economically undeveloped region, the number of students who are enriched with learning media available is an important factor that affects online learning behavior and learning outcomes.

Keywords

Online Learning, K-Prototypes Clustering, Economically Developed Region, Data Analysis, Different Groups, Learning Behavior, Learning Media

1. Introduction

Online learning has been growing continuously in the past two decades. Distance education has evolved from offline to online settings with the access to the Internet and affected by the COVID-19 epidemic, online learning has become a global learning method. Schools and universities have witnessed the unprecedented use of online collaboration tools and applications to support the continuing education of students and educators. The scale of digitally supported online learning or remote education increased exponentially in 2020, for those who can use digital devices connected to the Internet and likely changed the way education is provided forever [1]. There is no doubt that online education will become an important part of the new global education landscape.

During the COVID-19 epidemic, China carried out large-scale online learning. The purpose of this study is to research and analyze some of the learning problems of students. For this reason, we designed a questionnaire and collected a lot of relevant data. We introduced our research questions in detail in Section 1.2, and mainly introduced our questionnaire design and data collection and processing in Section 2.

1.1. Online Learning

Online learning has been on the increase in the past few decades. But many researchers have focused on specific areas of online education such as innovations in online learning strategies [2], quality in online education [3], designing sociable online learning environments [4], self-regulated learning in Open Online Courses [5], challenges of online learning [6], self-efficacy and self-regulation in online learning [7], attrition and achievement gaps in online learning [8], and online course dropout [9].

[10] reviewed research on online learning from 1993 to 2004. They reviewed 76 articles and divided the research into four themes: 1) course environment; 2) learners' outcomes; 3) learners' characteristics; and 4) institutional and administrative factors. The author describes the first theme as the course environment (n = 41, 53.9%) is an overarching theme that includes classroom culture, structural assistance, success factors, online interaction, and evaluation. [10] for their second theme found that studies focused on exploring the learning outcomes in the cognitive and affective domains through various research methods that have been used in the teaching process (n = 29, 38.2%). Another research theme focused on learners' characteristics (n = 12, 15.8%) and the social interaction, instructional design, and demographics of online learners. The final theme of their report was the institutional and administrative aspects (n = 13, 17.1%) in online learning. Their findings revealed that there was a lack of scholarly research in this area and most institutions did not have formal policies in place for course development as well as faculty and student support in training and evaluation [11].

[12] reviewed 695 articles on distance education and online learning from 2000 to 2008. In this review, the top three topics were interaction and communities of learning (n = 122, 17.6%), instructional design (n = 121, 17.4%) and learner

characteristics (*n* = 113, 16.3%).The least number of studies (less than 3%) found in studies examining the following research themes were these themes: management and organization (*n* = 18), research methods in DE and knowledge transfer (*n* = 13), globalization of education and cross-cultural aspects (*n* = 13), innovation and change (*n* = 13), and costs and benefits (*n* = 12). This study examined research areas in online learning, trends, priority areas, and gaps in distance education research.

[11] based on the previous systematic reviews [10] [12] [13], reviewed 619 articles on online learning from 2009 to 2018. Online learning research in this study is grouped into twelve different research themes which include Learner characteristics, Instructor characteristics, Course or program design and development, Course Facilitation, Engagement, Course Assessment, Course Technologies, Access, Culture, Equity, Inclusion, and Ethics, Leadership, Policy and Management, Instructor and Learner Support, and Learner Outcomes. In this review, the specific themes of Engagement (*n* = 179, 28.92%) and Learner Characteristics (*n* = 134, 21.65%) were the two topics that researchers like to study most. Articles focusing on Instructor Characteristics (*n* = 21, 3.39%) were least common in their statistics.

Table 1 shows some of the most and least researched themes on online learning in recent years. Current research in online learning is predominately focused on engagement and learner characteristics. Engagement themes can be subdivided into many areas, such as social presence [14] [15] [16], teaching presence [17] [18] [19], learner-learner interactions [15] [20] [21], participation patterns in online discussion [22] [23] and so on. Although many studies have been conducted on specific online learning topics, there are three problems with these studies: 1) it pays more attention to the system research of education and neglects the detailed teaching experience; 2) it is almost difficult to collect a large amount of data to analyze the research object; 3) there are few studies on the amount of learning media and the impact of learning behavior on students' learning effects. The content of our research solves these three deficiencies.

Table 1. Topics of previous online learning research work.

	1993-2004 (Tallent-Runnels [10])	2000-2008 (Zawacki [12])	2009-2018 (Martin [11])
Most number of studies	•Course environment •Learner outcomes	•Interaction and communities of learning •Instructional design •Learner characteristics	•Engagement •Learner characteristics •Evaluation and quality assurance
Lowest number of studies	•Learner characteristics •Institutional and administrative factors	•Management and organization •Research methods in DE and knowledge transfer •Globalization of education and cross-cultural aspects	•Course design and development •Leadership, policy, and management •Instructor characteristics

1.2. The Present Study

Based on questionnaires designed by education experts and related data sets collected, this study explores what are the effects of different educational resources on students' learning behaviors and online learning results. Lee [24] examined perceptions of adequate resources that could facilitate or inhibited students' adoption of an online learning system. They indicated that improvement of resources is necessary to help students to understand and use the online learning system. The study of [25] contributes to knowledge about how textbook resources could be leveraged in a bite-sized e-learning environment. Here we explore the difference between the hardware learning media of students in different clusters.

Finally, previous work has shown that perceived resources [24] have impacts on online learning adoption. The richer the perceived resources, the more positive influence on online learning. Accordingly, we will investigate how clusters with different hardware resources are different in online learning and examine the difference between students' learning media and their learning effects in economically developed regions and students in economically undeveloped regions. As such, this study was guided by four research questions:

- **Research question 1**: How to distinguish online learners through analyzing their questionnaire data, *i.e.*, how to separate similar online learners from dissimilar ones?
- **Research question 2:** What do the similar online learners have in common?
- **Research question 3**: How the online learners' learning behaviors, e.g., participation and learning time, are affected by learning media?
- **Research question 4**: What are the impacts of learning media on online learners' experiences, such as learning satisfaction and learning outcomes?

2. Methods

In order to study our problem, we designed a questionnaire and collected a large amount of data, then processed the data and used unsupervised machine learning method k-prototypes to cluster, too and statistics the data. Finally, the test method is used to test the hypothesis on the cluster data. The third part of our article is mainly about data processing and clustering, the fourth part is a statistical analysis of the data, and the fifth part is a hypothesis test, and some conclusions are very meaningful for the development of online education.

2.1. Study Design

This study is the latest data analysis using an unsupervised machine learning approach. The data for this study were from the questionnaire we collected. There are four main subjects of our survey, namely students, teachers, parents and school administrators. We invited education experts to design different questionnaires based on different roles. In our research, we mainly study the questionnaire of secondary school students. Our questionnaire has a total of 20 questions, including the region of the student, the grade the student is attending, the length of study time per day, learning behavior, learning status, learning expectations, etc.

2.1.1. Participants

Study participants were primary and secondary school students from an anonymous province in China, their parents, teachers, and their school administrators. All participants joined our study by filling out online questionnaires anonymously and voluntarily. The area where the students are located is distributed in both urban and rural areas, and the grade distribution is from primary school to high school.

The choice of participants using the online questionnaire can be found in the literature [26]. The total number of students in anonymous provinces in China is about 15 million. In China, students in grades 1 - 6, 7 - 9, and 10 - 12 are called primary school, middle school, and high school students, respectively. Approximately 37.5% students, their parents, teachers, and school administrators participated in the survey. All the people who participated in the questionnaire were viewed as ideal for this study as we were interested in what is the learning situation of students with different learning media in economically developed regions and economically undeveloped regions. This is the first large-scale online learning in an off-campus regular school, which provides important data for our study.

2.1.2. Collecting Data

We collected a total of 5,791,860 student questionnaires. Other common concerns with data we collected include potential cheating or speeding, where we define cheating as the inconsistency of the information before and after filling in the questionnaire, and we define speeding as clicking through questionnaire tasks as quickly as possible, paying minimal to no attention to the task itself. To eliminate the influence of these two factors, we added questions with consistent information to the questionnaire and recorded the time taken by the participants to fill out the questionnaire.

2.2. Data Analysis Procedures

2.2.1. Data Cleaning

Data cleaning techniques have been extensively covered in multiple surveys [27] [28] and tutorials [29] [30]. In our study, we mainly focus on student data, we defined our dirty data into the following three categories:

- **Data entry errors**: In our questionnaire, there is a question about how long do you study online every day, the time selection range we give is 0 - 15 hours, beyond this range is considered to be a wrong input data.
- **Cheating data errors**: In the questionnaire we designed, there are two questions, one of which is what is the location and category of your school, and the other is your grade. Based on the content of the first question, we can determine whether the answer to the second question filled in by the participant is correct. For example, the participant's first question answered is an urban secondary school,

so his grade should be between grade 7 and grade 12 (In our study, the first to sixth grades are defined as primary school, and the seventh to twelfth grades are defined as secondary school). If his grade is not in this specified range, we think his data is cheating data. The same method is used for primary school data.

- **"Speeding" data errors**: We recorded the time it took for each participant to fill in the questionnaire. The questionnaire we designed for students has a total of 20 questions and 90 options, so it takes at least 90 seconds to complete the questionnaire. If the time required for the participant to fill out the questionnaire is less than 90 seconds, then we regard it as speeding data.

According to the above three standards, we cleaned up student data, and the data of 775,516 participants were deleted. After the data was cleaned, there were still 5,015,344 participants' data. We have also cleaned up some redundant data options. After cleaning up, there are still 58 options.

2.2.2. Data Analysis

This study used a combination of unsupervised machine learning (k-prototypes clustering) and non-parametric statistical analyses. The objective of clustering is to partition a set of data objects into clusters such that data objects in the same cluster are more similar to each other than those in other clusters [31]. Partition clustering algorithms are widely applied clustering statistical methods. The k-means algorithm is used to analyze numeric data and the k-modes algorithm extends the k-means to cluster categorical data [32]. The k-prototypes algorithm integrates both the k-means and k-modes algorithms, to cluster mixed data [33]. K-prototypes clustering has been used education fields, such as virtual learning environment [34], educational contents [35], and with student health monitoring system [36] and other educational technologies [37]. For an overview of clustering analysis, see [38], and for clustering specifically in educational technology applications, see [39].

We perform k-prototypes clustering using the k modes [40] [41], pandas, and metrics packages in python. For k-prototypes clustering, one must determine how many clusters the analysis will create. Here, we use the Sum of Squared Errors (SSE) Score and Average Silhouette to determine the optimal number of clusters, which sometimes are referred to cluster validity metrics. We can invoke these two methods from the metrics package under scikit-learn in python language. Scikit-learn is a Python module for machine learning built on top of SciPy and is distributed under the 3-Clause BSD license. It provides various tools for model fitting, data preprocessing, model selection, model evaluation, and many other utilities. These methods are implemented according to the ideas of [42]. We used Euclidean distance metric when calculating the clusters. Because the data in our questionnaire has both numerical data and categorical data, we used the comprehensive evaluation method of Euclidean distance and Hamming distance to calculate clusters.

We also used a variety of non-parametric statistics due to non-normal data distributions after clustering. All non-parametric analyses were conducted by using IBM SPSS 23. SPSS Statistics software offers a range of advanced features, including ad hoc analysis, hypothesis testing and reporting. This makes it easier to access and manage data, select and perform analyses. We used Mann-Whitney-U-tests or Kruskal-Wallace-H tests, and for pairwise comparisons we report p-values adjusted with Bonferroni corrections.

3. Results

- **Research question 1**: How to distinguish online learners through analyzing their questionnaire data, *i.e.*, how to separate similar online learners from dissimilar ones?

We use the unsupervised clustering method k-prototypes to distinguish learners with different characteristics. Before using the k-prototypes clustering method to cluster the data, we first used Sum of Squared Errors and average silhouette to evaluate how the data should be clustered into several categories. The two effects of evaluating the optimal number of clusters are shown in the Figure 1 and Figure 2 respectively. The following pictures respectively show the two effects of evaluating the optimal number of clusters. Figure 1 is the evaluation result of the sum of squared errors score method, and Figure 2 is the evaluation result of the average silhouette method. For picture 1, the k value at the inflection point where the score drops faster is the number of groups that should be divided. For Figure 2, the k value corresponding to the point with the highest score is the number of people that should be divided. According to the evaluation results, when the number of clusters k is 2, the data classification effect is the best. Therefore, we use k = 2 to cluster online learners in developed and undeveloped regions.

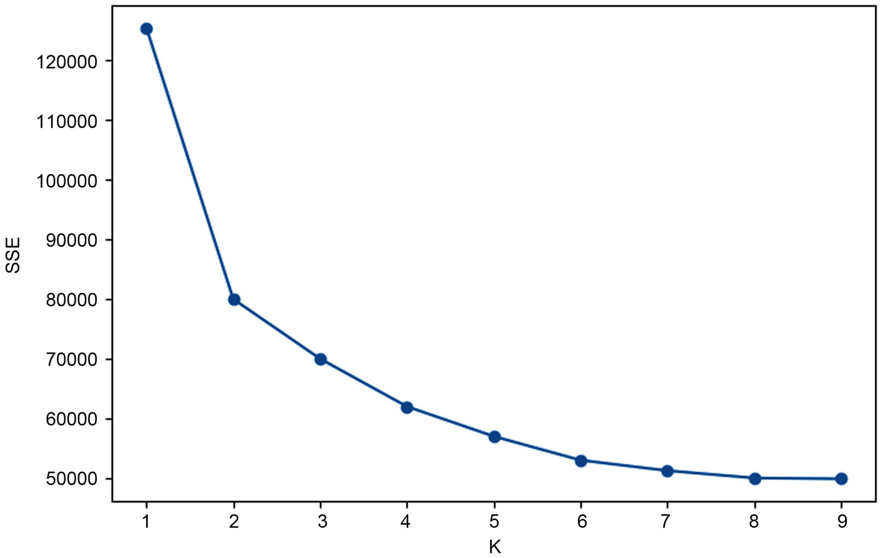

Figure 1. Sum of squared errors score (SSE).

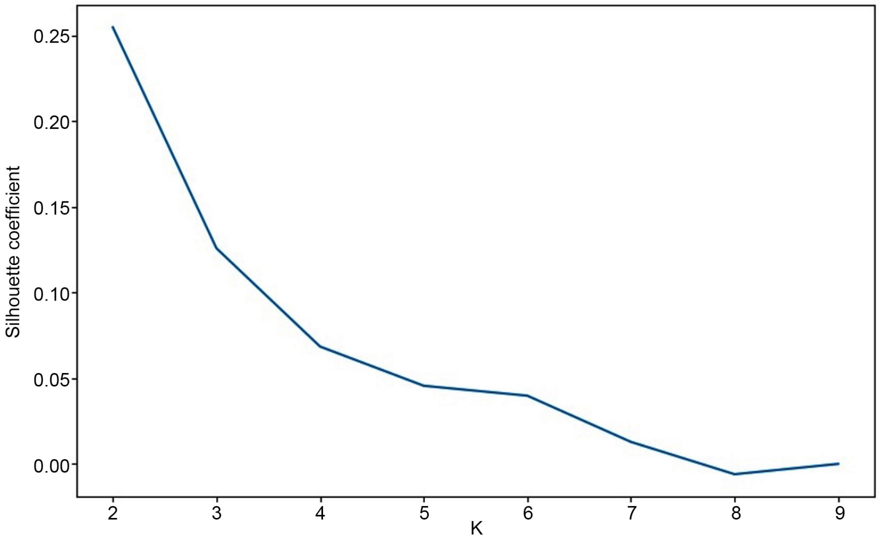

Figure 2. Average silhouette score (silhouette coefficient).

• **Research question 2**: What do the similar online learners have in common, and what are the characteristics between dissimilar online learners?

According to the question types designed by our survey data questionnaire, we divide the questions into three major questions, which are our question 2 to question 4. Question 2 is mainly the objective information part of the questionnaire filled out by online learners. Here we include the following questions:

- RQ2-1: Your grade?
- RQ2-2: Which of the following equipment/materials did you use during your online learning throughout the outbreak?
- RQ2-3: What is the longest class time for your online courses?
- RQ2-4: How long do you study online every day (move the slider to the corresponding number of positions, 0 - 15 hours)?
- RQ2-5: What content does your online course include?

We attribute these questions to objective questions. Questions 3 and 4 mainly discuss issues related to students' subjective wishes. In later chapters, we will detail the content of their questions.

According to the estimated k value, we divide the undeveloped regions into two clusters, and the economically developed regions are also divided into two clusters. We have made statistics on the results divided into two clusters. Table 2 is the statistical results of the two clusters in the undeveloped regions, and Table 3 is the statistical results of the two clusters in the developed regions. In the table, we have counted the results of the ten options contained in the above questions. The main indicators included are the minimum (Min), maximum (Max), average (Mean), standard deviation (SD), and percentage (Percent) of the cluster.

Let's first compare the data of two clusters in undeveloped regions. In Table 1, we can clearly see that cluster 1 and cluster 2 are significantly different in the distribution of these ten options. In terms of the study online hours, the average learning time of cluster 1 is 4.75, while the average learning time of cluster 2 is 10.16. In terms of learning media used, cluster 1 and cluster 2 mainly use smartphones for online learning, but cluster 1 and cluster 2 have some gaps in computers, tablets, and paper materials. The percentages of cluster 2 for these three options are higher than cluster 1, for example, cluster 1 is 26.5% for online learning using computers, and cluster 2 is 30.4%. Computers, tablets, and other learning equipment can only be purchased under certain economic conditions, which indicate that the second group may be the group with better learning media in economically undeveloped regions. At the same time, the percentage of special education in cluster 2 is higher than that in cluster 1, which indicates that the educational resources of the school in cluster 2 should be better than that in cluster 1. These results indicate that clusters with richer teaching and learning media may have better learning behaviors and learning effects.

Table 2. Cluster data statistics of undeveloped regions.

	Cluster 1 (n = 695,679)					Cluster 2 (n = 340,808)				
	Min	Max	Mean	SD	Percent (%)	Min	Max	Mean	SD	Percent (%)
RQ2-1	7	12	8.99	1.614	–	7	12	9.18	1.663	–
RQ2-2	1	4	3.07	0.881	–	1	4	3.33	0.768	–
RQ2-3	0	8	4.75	1.641	–	7	15	10.16	2.253	–
RQ2-4	0	1	0.26	0.441	26.5	0	1	0.30	0.460	30.4
RQ2-5	0	1	0.11	0.318	11.4	0	1	0.14	0.346	13.9
RQ2-6	0	1	0.86	0.347	86	0	1	0.84	0.362	84.5
RQ2-7	0	1	0.33	0.468	32.50	0	1	0.39	0.487	38.6
RQ2-8	0	1	0.89	0.317	88.70	0	1	0.88	0.330	87.6
RQ2-9	0	1	0.77	0.418	77.50	0	1	0.83	0.371	3.5
RQ2-10	0	1	0.47	0.499	47.10	0	1	0.55	0.498	54.6

Table 3. Cluster data statistics of developed regions.

	Cluster 1 (n = 623,367)					Cluster 2 (n = 413,120)				
	Min	Max	Mean	SD	Percent (%)	Min	Max	Mean	SD	Percent (%)
RQ2-1	7	12	8.97	1.631	–	7	12	9.17	1.648	–
RQ2-2	1	4	3.09	0.855	–	1	4	3.37	0.728	–
RQ2-3	0	9	5.15	1.650	–	7	15	10.17	2.249	–
RQ2-4	0	1	0.41	0.493	41.4	0	1	0.53	0.499	52.8
RQ2-5	0	1	0.19	0.395	19.4	0	1	0.24	0.427	24.0
RQ2-6	0	1	0.86	0.347	86	0	1	0.84	0.362	84.5
RQ2-7	0	1	0.33	0.468	32.50	0	1	0.39	0.487	38.6
RQ2-8	0	1	0.89	0.317	88.70	0	1	0.88	0.330	87.6
RQ2-9	0	1	0.77	0.418	77.50	0	1	0.83	0.371	83.5
RQ2-10	0	1	0.47	0.499	47.1	0	1	0.55	0.498	54.6

For the two clusters in the economically developed regions, the same phenomenon exists between them as in the economically undeveloped regions. The teaching and learning media of cluster 2 are more abundant than the learning media of cluster 1. However, the gap between cluster 2 and cluster 1 in economically developed regions is greater than that in economically undeveloped regions. For example, in economically undeveloped regions, the average gap between cluster 2 and cluster 1 that uses computers for online learning is 11.4%, while the average gap between clusters 2 of computers for online learning using computers in economically undeveloped regions is 3.9%. We can also find this result by comparing the standard deviation of learning media and teaching resources in economically developed regions and economically undeveloped regions, for example cluster 2 developed (computer SD = 0.499) > cluster 2 undeveloped (computer SD = 0.460).

Comparing the clusters of economically developed regions and economically undeveloped regions, we found some interesting results. We compare cluster 1 and cluster 2 in developed regions and undeveloped regions respectively, we can find cluster 1 developed (computer Mean = 0.41) > cluster 1 undeveloped (computer Mean = 0.26), cluster 2 developed (computer Mean = 0.53) > cluster 2 undeveloped (computer Mean = 0.30), cluster 1 developed (Tablet Mean = 0.19) > cluster 1 undeveloped (Tablet Mean = 0.11), cluster 2 developed (Tablet Mean = 0.24) > cluster 2 undeveloped (Tablet Mean = 0.14). This shows that there are not only differences in learning media within the same region, but also differences in learning media between different regions. However, although the differences between different clusters in developed regions are greater than the differences between different clusters in undeveloped regions, developed regions generally have more learning media than undeveloped regions. And the average online learning hours of students in two clusters in developed regions are higher than the average online learning hours of students in undeveloped regions. The average online learning time of students in the two clusters in developed regions is higher than that of students in undeveloped regions, which indicates that students with different teacher resources and learning media may have different effects and feelings in online learning.

Research question 3: How the online learners' learning behaviors, e. G., participation and learning time, are affected by learning media?

We divide students' online learning behaviors into three parts: students' classroom learning behaviors, students' learning behaviors when they encounter problems, and students' learning behaviors after class. The details of these three parts are as follows:

- RQ3-1: Homework submission;
- RQ3-2: In-class test;
- RQ3-3: Video conference;
- RQ3-4: In-class commenting;
- RQ3-5: Viewing homework that achieved an excellent grade;

- RQ3-6: Screen sharing;
- RQ3-7: Live commenting;
- RQ3-8: Discussion;
- RQ3-9: Solving independently by searching online;
- RQ3-10: Re-watch recorded lectures when you encounter knowledge points that you have not mastered;
- RQ3-11: Attend Q&A sessions organized by teachers;
- RQ3-12: Ask teachers by using social platforms;
- RQ3-13: Communicate with other students;
- RQ3-14: Re-watch lecture videos after class;
- RQ3-15: Carefully studied other course materials provided by your teacher;
- RQ3-16: Carried out home-based self-study activities;
- RQ3-17: Ask the teacher when you encounter a problem;
- RQ3-18: The quality of the work done online be as good as offline.

Table 4 and Table 5 are statistics on the learning behaviors of students in undeveloped and developed regions respectively. Items 1 - 8 in the table are students' classroom learning behaviors, and items 9 - 13 are students' learning behaviors when they encounter problems, 14 - 18 items are students' learning behaviors after class. These three types of learning behaviors correspond to questions 4, 11, and 14 in our questionnaire. Table 4 is the statistics of online learning behaviors of students in different clusters in undeveloped regions. By observing the table data, we can see that cluster 2 performs better than cluster 1 in these three types of learning behaviors. In terms of classroom learning behaviors, the most common behaviors for students are homework submission, class testing and excellent homework viewing. Among them, more than 80% of each cluster has submitted homework, and the least common behavior is screen sharing, only about 10%. There is a big gap between cluster 1 and cluster 2 in classroom test, cluster 2 undeveloped (In-class-test Mean = 0.44) > cluster 1 undeveloped (In-class-test Mean = 0.33), When students encounter knowledge they don't understand in online learning, they usually re-watch the recorded lectures or independently searching online to solve the problem. They seldom participate in the Q&A sessions organized by teachers or use social platforms to ask teachers. However, we can still find that the data of students in cluster 2 is better than the data of students in cluster 1 in terms of participating in the question and answer session of the teacher organization and using the social platform to ask the teacher, cluster 2 undeveloped (Q&A Mean = 0.34) > cluster 1 undeveloped (Q&A Mean = 0.25), cluster 2 undeveloped (Ask teachers by using social platforms Mean = 0.34) > cluster 1 undeveloped (Ask teachers by using social platforms Mean = 0.25) This shows that cluster 2 may have better teacher resources than cluster 1. After the class, most of the students can earnestly study other course materials provided by the teacher and carry out self-study activities at home.

Table 5 is the statistical information of different clusters in developed regions. In developed regions, we can find that cluster 2 performs better than cluster 1,

Table 4. Cluster data statistics of undeveloped regions.

	Cluster 1 (n = 695,679)					Cluster 2 (n = 340,808)				
	Min	Max	Mean	SD	Percent (%)	Min	Max	Mean	SD	Percent (%)
RQ3-1	0	1	0.83	0.377	82.9	0	1	0.87	0.340	86.6
RQ3-2	0	1	0.33	0.471	33.2	0	1	0.44	0.496	44.0
RQ3-3	0	1	0.19	0.390	18.7	0	1	0.27	0.443	26.8
RQ3-4	0	1	0.34	0.474	34.1	0	1	0.42	0.493	41.8
RQ3-5	0	1	0.41	0.492	41.3	0	1	0.49	0.500	48.6
RQ3-6	0	1	0.10	0.302	10.1	0	1	0.14	0.350	14.3
RQ3-7	0	1	0.19	0.396	19.5	0	1	0.22	0.411	21.5
RQ3-8	0	1	0.25	0.435	25.3	0	1	0.31	0.462	31.0
RQ3-9	0	1	0.57	0.495	56.9	0	1	0.61	0.487	61.3
RQ3-10	0	1	0.70	0.457	70.3	0	1	0.72	0.447	72.4
RQ3-11	0	1	0.25	0.433	25.0	0	1	0.34	0.472	33.6
RQ3-12	0	1	0.34	0.475	34.3	0	1	0.44	0.496	43.5
RQ3-13	0	1	0.45	0.498	45.2	0	1	0.51	0.500	51.2
RQ3-14	0	1	0.85	0.362	84.5	0	1	0.87	0.335	87.1
RQ3-15	0	1	0.88	0.324	88.1	0	1	0.92	0.266	92.4
RQ3-16	0	1	0.79	0.410	78.6	0	1	0.85	0.356	85.1
RQ3-17	0	1	0.53	0.499	52.7	0	1	0.61	0.487	61.3
RQ3-18	0	1	0.64	0.480	64.0	0	1	0.68	0.466	68.2

Table 5. Cluster data statistics of developed regions.

	Cluster 1 (n = 623,367)					Cluster 2 (n = 413,120)				
	Min	Max	Mean	SD	Percent (%)	Min	Max	Mean	SD	Percent (%)
RQ3-1	0	1	0.87	0.334	87.2	0	1	0.91	0.287	91.0
RQ3-2	0	1	0.44	0.855	44.4	0	1	0.60	0.490	59.7
RQ3-3	0	1	0.31	0.463	31.2	0	1	0.42	0.494	42.1
RQ3-4	0	1	0.51	0.500	51.5	0	1	0.61	0.488	60.9
RQ3-5	0	1	0.54	0.498	54.2	0	1	0.63	0.483	62.8
RQ3-6	0	1	0.16	0.370	16.3	0	1	0.23	0.421	23.0
RQ3-7	0	1	0.27	0.442	26.5	0	1	0.31	0.460	30.5
RQ3-8	0	1	0.34	0.474	34.0	0	1	0.42	0.495	41.7
RQ3-9	0	1	0.61	0.489	60.5	0	1	0.66	0.475	65.5
RQ3-10	0	1	0.72	0.450	71.9	0	1	0.74	0.436	74.5
RQ3-11	0	1	0.34	0.473	33.8	0	1	0.44	0.497	44.3
RQ3-12	0	1	0.44	0.496	43.6	0	1	0.57	0.496	56.5
RQ3-13	0	1	0.52	0.500	51.7	0	1	0.59	0.492	59.1
RQ3-14	0	1	0.84	0.363	84.4	0	1	0.87	0.339	86.7
RQ3-15	0	1	0.91	0.280	91.4	0	1	0.94	0.239	93.9
RQ3-16	0	1	0.84	0.370	83.6	0	1	0.88	0.321	88.3
RQ3-17	0	1	0.61	0.489	60.6	0	1	0.68	0.465	68.5
RQ3-18	0	1	0.70	0.458	69.9	0	1	0.73	0.446	72.7

and the gap in some aspects is relatively large. For example, the proportion of students in cluster 2 taking classroom exams is 16% higher than that of students in cluster 1, and the proportions of students in cluster 2 participating in teacher-organized question-and-answer sessions and using social platforms to ask teacher questions are 10% and 13% higher than those in cluster 1, respectively. This shows that even in developed regions, due to differences in learning media, different student groups may have different learning behaviors. Obviously, in terms of various learning behaviors, the students in cluster 1 are worse than the students in cluster 2, and the gap is larger in developed regions.

Although due to differences in teacher resources/learning media, developed regions and undeveloped regions have divided into different student groups, according to the data in **Table 4** and **Table 5**, we find that the learning behaviors of students in developed regions are generally better than those in undeveloped regions. We compare cluster 1 in developed regions with cluster 1 in undeveloped regions and compare cluster 2 in developed regions with cluster 2 in undeveloped regions. We want to use this comparison to find the differences between the groups with fewer learning media in developed and undeveloped regions through this comparison, and the differences between the groups with more teacher resources/learning media. First, comparing cluster 1, we can find that students in cluster 1 in developed regions are significantly more prominent in learning behaviors, such as In-class-test (cluster 1 developed (Mean = 0.44) > cluster 1 undeveloped (Mean = 0.33)), video conferences (cluster 1 developed (Mean = 0.31) > cluster 1 undeveloped (Mean = 0.19)), In-class commenting (cluster 1 developed (Mean = 0.51) > cluster 1 undeveloped (Mean = 0.34)), and ask teachers through social platforms (cluster 1 developed (Mean = 0.44) > cluster 1 undeveloped (Mean = 0.34)), than students in cluster 1 in undeveloped regions. In terms of other learning behaviors, there is a gap between cluster 1 in undeveloped regions and cluster 1 in developed regions, but the gap is not large. Secondly, comparing cluster 2 students in different regions, developed regions performed significantly better than undeveloped regions in learning behaviors such as In-class test, video conferences, in class commenting, viewing homework that achieved an excellent grade, and attend Q&A sessions organized by teachers. This is a comparison between the groups with better learning media in developed and undeveloped regions, but there is a big difference between them, which shows that the economic development status determines the difference in learning media, which will be to a large extent affect students' learning behavior.

- **Research question 4**: What are the impacts of learning media on online learners' experiences, such as learning satisfactory and learning outcomes?

The detailed content of the question can be divided into the following 24 items:

- RQ4-1: Your learning statuses;
- RQ4-2: Like webcast;
- RQ4-3: Like recording;

- RQ4-4: Like resource pack;
- RQ4-5: Frequency of classroom interaction in online learning;
- RQ4-6: Poor experience with online learning platforms;
- RQ4-7: Insufficient communication with teachers;
- RQ4-8: Eyestrain caused by long staring at screens;
- RQ4-9: Confusion in setting up the platforms;
- RQ4-10: Self-learning ability;
- RQ4-11: Utilization of digital resources ability;
- RQ4-12: Communication ability;
- RQ4-13: Webcast satisfaction;
- RQ4-14: Lecture recording satisfaction;
- RQ4-15: Teachers' attitude satisfaction;
- RQ4-16: Online learning media satisfaction;
- RQ4-17: Access to courses delivered by famous teachers;
- RQ4-18: More convenient to review course content;
- RQ4-19: Achieve better learning performance;
- RQ4-20: Can learn anytime and anywhere;
- RQ4-21: Less effective than classroom-based education;
- RQ4-22: Unstable course quality;
- RQ4-23: Increased learning efforts;
- RQ4-24: Lack of teacher-student interaction.

In addition to differences in teacher resources, learning media, and learning behaviors for students online learning, there are also some differences in the learning experience. Table 6 and Table 7 are the data of developed and undeveloped regions that we obtained after clustering the data. Table 6 is the result of statistical data after the undeveloped regions are divided into two clusters. Comparing the data of cluster 1 and cluster 2 in undeveloped regions, there are some obvious differences. Comparing the data of cluster 1 and cluster 2 in undeveloped regions, there are some obvious differences. The students in cluster 1 are not as serious as the students in cluster 1 (cluster 1 undeveloped (Mean = 2.46) > cluster 2 developed (Mean = 2.30)), and the probability of eyestrain caused by long staring at screens in cluster 2 is higher than that in cluster 1 (cluster 1 undeveloped (Mean = 0.75) < cluster 2 undeveloped (Mean = 0.81)). At the same time, Cluster 2 also has a high probability in terms of confusion in setting up the platforms (cluster 1 undeveloped (Mean = 0.17) < cluster 2 undeveloped (Mean = 0.22)) and digital resource utilization capabilities (cluster 1 undeveloped (Mean = 0.34) < cluster 2 undeveloped (Mean = 0.38)). This indicates that the students in cluster 2 have more online learning media than those in cluster 1. The students in cluster 2 use more teaching software and believe that they have cultivated their digital resource utilization ability. This result can also be found by analyzing their satisfaction with teachers (cluster 1 undeveloped (Mean = 1.92) > cluster 2 undeveloped (Mean = 1.90)) and online learning media (cluster 1 undeveloped (Mean = 2.13) > cluster 2 undeveloped (Mean = 2.09)). Students in cluster 2 are more satisfied with teachers and learning media.

Table 7 shows the statistical results of two clusters in developed regions. The result of the difference between the two clusters in the developed regions is basically the same as the result of the difference between the two clusters in the undeveloped regions, but the differences between some internal clusters in the developed regions may be greater, for example, digital utilization ability (cluster 1 developed (Mean = 0.42) < cluster 2 developed (Mean = 0.53)), communication ability (cluster 1 developed (Mean = 0.26) < cluster 2 developed (Mean = 0.33)).

To compare the difference between developed regions and undeveloped regions, we also compare the developed regions cluster 1 and the undeveloped regions cluster 1, and the developed regions cluster 2 and the undeveloped regions cluster 2 are compared. For cluster 1, groups in developed regions have better experience in learning status (cluster 1 undeveloped (Mean = 2.46) > cluster 1 developed (Mean = 2.31)), satisfaction with teachers (cluster 1 undeveloped (Mean = 1.92) > cluster 1 developed (Mean = 1.79)), satisfaction with online learning

Table 6. Cluster data statistics of undeveloped regions.

	Cluster 1 (n = 695,679)					Cluster 2 (n = 340,808)				
	Min	Max	Mean	SD	Percent (%)	Min	Max	Mean	SD	Percent (%)
RQ4-1	1	5	2.46	1.187	–	1	5	2.30	1.170	–
RQ4-2	0	1	0.52	0.500	51.9	0	1	0.55	0.497	55.4
RQ4-3	0	1	0.24	0.428	24.2	0	1	0.25	0.432	24.8
RQ4-4	0	1	0.15	0.358	15.1	0	1	0.17	0.376	17.1
RQ4-5	1	5	2.59	1.191	–	1	5	2.64	1.099	–
RQ4-6	0	1	0.23	0.419	22.7	0	1	0.24	0.430	24.5
RQ4-7	0	1	0.22	0.417	22.4	0	1	0.24	0.435	23.6
RQ4-8	0	1	0.75	0.431	75.4	0	1	0.81	0.394	80.7
RQ4-9	0	1	0.17	0.379	17.4	0	1	0.22	0.414	21.9
RQ4-10	0	1	0.78	0.411	78.5	0	1	0.82	0.385	81.9
RQ4-11	0	1	0.34	0.472	33.6	0	1	0.38	0.486	38.3
RQ4-12	0	1	0.23	0.418	22.6	0	1	0.28	0.447	27.6
RQ4-13	1	4	2.26	0.801	–	1	4	2.22	0.821	–
RQ4-14	1	4	2.24	0.797	–	1	4	2.20	0.818	–
RQ4-15	1	4	1.92	0.752	–	1	4	1.90	0.772	–
RQ4-16	1	4	2.13	0.757	–	1	4	2.09	0.776	–
RQ4-17	0	1	0.39	0.488	39.1	0	1	0.44	0.497	44.1
RQ4-18	0	1	0.79	0.410	78.6	0	1	0.81	0.396	80.6
RQ4-19	0	1	0.15	0.352	14.5	0	1	0.19	0.392	19.0
RQ4-20	0	1	0.56	0.497	55.8	0	1	0.56	0.496	56.5
RQ4-21	0	1	0.67	0.468	67.5	0	1	0.66	0.472	66.4
RQ4-22	0	1	0.29	0.454	29.1	0	1	0.32	0.466	31.8
RQ4-23	0	1	0.21	0.410	21.3	0	1	0.28	0.448	27.9
RQ4-24	0	1	0.59	0.491	59.2	0	1	0.58	0.493	58.2

Table 7. Cluster data statistics of developed regions.

	Cluster 1 (n = 623,367)					Cluster 2 (n = 413,120)				
	Min	Max	Mean	SD	Percent (%)	Min	Max	Mean	SD	Percent (%)
RQ4-1	1	5	2.30	1.136	–	1	5	2.21	1.114	–
RQ4-2	0	1	0.59	0.492	58.7	0	1	0.62	0.486	61.7
RQ4-3	0	1	0.22	0.414	21.9	0	1	0.24	0.430	24.4
RQ4-4	0	1	0.16	0.363	15.6	0	1	0.19	0.389	18.5
RQ4-5	1	5	2.51	0.981	–	1	5	2.61	0.942	–
RQ4-6	0	1	0.22	0.411	21.5	0	1	0.25	0.430	24.5
RQ4-7	0	1	0.22	0.413	21.9	0	1	0.24	0.429	24.4
RQ4-8	0	1	0.77	0.423	76.6	0	1	0.83	0.376	82.9
RQ4-9	0	1	0.18	0.384	18.0	0	1	0.22	0.417	22.4
RQ4-10	0	1	0.81	0.392	81.0	0	1	0.83	0.371	83.5
RQ4-11	0	1	0.42	0.493	41.6	0	1	0.53	0.499	52.6
RQ4-12	0	1	0.26	0.440	26.2	0	1	0.33	0.470	32.9
RQ4-13	1	4	2.08	0.791	–	1	4	2.04	0.810	–
RQ4-14	1	4	2.10	0.800	–	1	4	2.06	0.823	–
RQ4-15	1	4	1.79	0.728	–	1	4	1.78	0.754	–
RQ4-16	1	4	1.98	0.751	–	1	4	1.94	0.772	–
RQ4-17	0	1	0.34	0.474	34.0	0	1	0.39	0.488	38.9
RQ4-18	0	1	0.81	0.393	81.0	0	1	0.84	0.371	83.6
RQ4-19	0	1	0.16	0.369	16.2	0	1	0.21	0.408	21.1
RQ4-20	0	1	0.55	0.497	55.2	0	1	0.58	0.493	58.2
RQ4-21	0	1	0.68	0.467	67.9	0	1	0.68	0.467	67.7
RQ4-22	0	1	0.28	0.451	28.4	0	1	0.32	0.468	32.4
RQ4-23	0	1	0.22	0.415	22.1	0	1	0.31	0.464	31.5
RQ4-24	0	1	0.56	0.496	56.4	0	1	0.56	0.496	56.3

media (cluster 1 undeveloped (Mean = 2.13) > cluster 1 developed (Mean = 1.98)), and the ability to use digital resources (cluster 1 undeveloped (Mean = 0.34) < cluster 1 developed (Mean = 0.42)) than those in undeveloped regions. On the contrary, compared with developed regions, groups in undeveloped regions lack teacher-student interaction (cluster 1 undeveloped (Mean = 0.59) > cluster 1 developed (Mean = 0.56)), and are better able to access courses delivered by famous teachers (cluster 1 undeveloped (Mean = 0.39) > cluster 1 developed (Mean = 0.34)). For cluster 2, the difference between developed and undeveloped regions is similar to that of cluster 1. This indicates that the difference in learning media will lead to differences in students' online learning experience. In general, there is more learning media in developed regions, and the higher their satisfaction with online learning, the better their experience. At the same time, they also feel more tired and use more online teaching software. Students in undeveloped regions are not as good as those in developed regions in terms of online learning satisfaction and learning fatigue, but they believe that online education has changed the inequality of educational resources to a certain extent, allowing them to hear more famous teacher courses.

4. Conclusions

In the previous sections, we counted the data when answering the questions. Although the statistics of different groups in different regions are different, are their distributions the same? To this end, we performed a Mann-Whitney-U-Test. We have done four types of Mann-Whitney-U-Test, namely, developed region cluster 1 and cluster 2, undeveloped region cluster 1 and cluster 2, developed region cluster 1 and undeveloped region cluster 1, as well as developed region cluster 2 and undeveloped region cluster 2. According to these four tests, we tested their performance in question 2 to question 4 respectively, and their P values were all at the level of 0.000** < 0.001, which indicates that the distributions between different groups are not the same. Therefore, our research method of collecting data from different groups and comparing them is very meaningful. Online education has always been a research direction of educators. Due to the COVID-19 epidemic, large-scale online learning by students provides a good case for our research. This allows us to collect large-scale data that is difficult to collect normally, because, under normal circumstances, there are not so many people who choose to study online. When most scholars research online learning, the participants are often relatively limited. For example, many online learning participants may be data from a certain website or data from a certain university. This has led to very little research on the online learning situation of students in different regions. Our research can provide good reference for future research on online learning and economic development. Education plays an important role in economic development, and at the same time, the economy provides an important guarantee for the development of education. Higher education has provided a positive and significant impact on economic development, and engineering and natural science majors have played the most prominent role in this process [43]. The research of Mishra and Agarwal [44] showed that for undeveloped countries, economic growth often corresponds to an increase in education expenditure. Our research is also based on the different conditions of economic development, to study the differences in the online learning status of different student groups. We asked four questions in total and answered them based on the results of statistical data. We can summarize the results of these four questions as follows:

1) Whether in developed or undeveloped regions, students have different group gathering effects based on their own learning media, teacher resources, and learning behaviors;

2) Within the same area, the differences between different groups are relatively large in terms of learning media. The difference between different groups in undeveloped regions is smaller than the difference between different groups in

developed regions;

3) Learning media in developed regions are better than those in undeveloped regions. This leads to students in developed regions that perform better than students in undeveloped regions in terms of online learning behavior, and the learning experience is better than that in undeveloped regions;

4) For students in developed regions, spend a long time on online learning on average, have more learning software installed, and the number of times teachers tutor students is relatively large, so they feel more exhausted than students in undeveloped regions;

5) For students in undeveloped regions, their study time, learning behavior, and good learning experience are not as good as in developed regions, but they can feel that online education has brought them better courses and let them hear the courses of famous teachers.

This is very helpful for us to carry out online learning in areas where the economic development of different learning media is unbalanced in the future. With the hardware supporting facilities, online learning can help us change the unbalanced state of teacher distribution.

However, our research also has some shortcomings. Our data did not have test scores, which makes it very difficult for us to analyze whether the performance of online learning students is better than offline students in the future. And we can also study whether students' good online learning behaviors will affect students' academic performance. We may try some more on this.

Funding Information

This paper and the study are funded by: Fund Project 1: NSFC (Project No.: 62077028, 61877029), Fund Project 2: the Science and Technology Planning Project of Guangdong (Project No.: 2021B0101420003, 2020B0909030005, 2020B1212030003, 2019A050510024, 2020ZDZX3013, 2019B1515120010, 2018-KTSCX016), Fund Project 3: the Science and Technology Planning Project of Guangzhou (Project No.: 201902010041), Fund Project 4: the teaching reform research projects of Jinan University (JG2021112) and The High Performance Public Computing Service Platform of Jinan University.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Li, C. and Lalani, F. (2020) The COVID-19 Pandemic Has Changed Education Forever. This Is How. World Economic Forum.

[2] Davis, D., *et al.* (2018) Activating Learning at Scale: A Review of Innovations in Online Learning Strategies. *Computers & Education*, **125**, 327-344. https://doi.org/10.1016/j.compedu.2018.05.019

[3] Esfijani, A. (2018) Measuring Quality in Online Education: A Meta-Synthesis. *American Journal of Distance Education*, **32**, 57-73. https://doi.org/10.1080/08923647.2018.1417658

[4] Weidlich, J. and Bastiaens, T.J. (2019) Designing Sociable Online Learning Environments and Enhancing Social Presence: An Affordance Enrichment Approach. *Computers & Education*, **142**, Article ID: 103622. https://doi.org/10.1016/j.compedu.2019.103622

[5] Jansen, R.S., *et al.* (2020) Supporting Learners' Self-Regulated Learning in Massive Open Online Courses. *Computers & Education*, **146**, Article ID: 103771. https://doi.org/10.1016/j.compedu.2019.103771

[6] Gillett-Swan, J. (2017) The Challenges of Online Learning: Supporting and Engaging the Isolated Learner. *Journal of Learning Design*, **10**, 20-30. https://doi.org/10.5204/jld.v9i3.293

[7] Bradley, R.L., *et al.* (2017) Examining the Influence of Self-Efficacy and Self-Regulation in Online Learning. *College Student Journal*, **51**, 518-530.

[8] Kizilcec, G., *et al.* (2015) Attrition and Achievement Gaps in Online Learning. *Proceedings of the Second* (2015) *ACM Conference on Learning*, Vancouver, 14-18 March 2015, 57-66. https://doi.org/10.1145/2724660.2724680

[9] de la Varre, C., *et al.* (2014) Reasons for Student Dropout in an Online Course in a Rural K-12 Setting. *Distance Education*, **35**, 324-344. https://doi.org/10.1080/01587919.2015.955259

[10] Tallent-Runnels, M.K., *et al.* (2006) Teaching Courses Online: A Review of the Research. *Review of Educational Research*, **76**, 93-135. https://doi.org/10.3102/00346543076001093

[11] Martin, F., Sun, T. and Westine, C.D. (2020) A Systematic Review of Research on Online Teaching and Learning from 2009 to 2018. *Computers & Education*, **159**, Article ID: 104009. https://doi.org/10.1016/j.compedu.2020.104009

[12] Zawacki-Richter, O., Baecker, E.M. and Vogt, S. (2009) Review of Distance Education Research (2000 to 2008): Analysis of Research Areas, Methods, and Authorship Patterns. *The International Review of Research in Open and Distributed Learning*, **10**, 21-50. https://doi.org/10.19173/irrodl.v10i6.741

[13] Hung, J.L. (2012) Trends of E-Learning Research from 2000 to 2008: Use of Text Mining and Bibliometrics. *British Journal of Educational Technology*, **43**, 5-16. https://doi.org/10.1111/j.1467-8535.2010.01144.x

[14] Akcaoglu, M. and Lee, E. (2016) Increasing Social Presence in Online Learning through Small Group Discussions. *The International Review of Research in Open and Distributed Learning*, **17**. https://doi.org/10.19173/irrodl.v17i3.2293

[15] Phirangee, K. (2016) Students' Perceptions of Learner-Learner Interactions that Weaken a Sense of Community in an Online Learning Environment. *Online Learning*, **20**, 13-33.

[16] Wei, C.W., Chen, N.S. and Kinshuk (2012) A Model for Social Presence in Online Classrooms. *Educational Technology Research and Development*, **60**, 529-545. https://doi.org/10.1007/s11423-012-9234-9

[17] Orcutt, J.M. and Dringus, L.P. (2017) Beyond Being There: Practices that Establish Presence, Engage Students and Influence Intellectual Curiosity in a Structured Online Learning Environment. *Online Learning*, **21**, 15-35.

[18] Preisman, K.A. (2014) Teaching Presence in Online Education: From the Instructor's Point of View. *Online Learning*, **18**, n3.

[19] Wisneski, J.E., Ozogul, G. and Bichelmeyer, B.A. (2015) Does Teaching Presence Transfer between MBA Teaching Environments? A Comparative Investigation of Instructional Design Practices Associated with Teaching Presence. *The Internet and Higher Education*, **25**, 18-27. https://doi.org/10.1016/j.iheduc.2014.11.001

[20] Tawfik, A.A., *et al.* (2018) Effects of Success v Failure Cases on Learner-Learner Interaction. *Computers & Education*, **118**, 120-132. https://doi.org/10.1016/j.compedu.2017.11.013

[21] Shackelford, J.L. and Maxwell, M. (2012) Sense of Community in Graduate Online Education: Contribution of Learner to Learner Interaction. *The International Review of Research in Open and Distributed Learning*, **13**, 228-249. https://doi.org/10.19173/irrodl.v13i4.1339

[22] Marbouti, F. and Wise, A.F. (2016) Starburst: A New Graphical Interface to Support Purposeful Attention to Others' Posts in Online Discussions. *Educational Technology Research and Development*, **64**, 87-113. https://doi.org/10.1007/s11423-015-9400-y

[23] Wise, A.F., *et al.* (2012) Microanalytic Case Studies of Individual Participation Patterns in an Asynchronous Online Discussion in an Undergraduate Blended Course. *The Internet and Higher Education*, **15**, 108-117. https://doi.org/10.1016/j.iheduc.2011.11.007

[24] Lee, Y.C. (2008) The Role of Perceived Resources in Online Learning Adoption. *Computers & Education*, **50**, 1423-1438. https://doi.org/10.1016/j.compedu.2007.01.001

[25] Lau, K.H., *et al.* (2018) The Role of Textbook Learning Media in E-Learning: A Taxonomic Study. *Computers & Education*, **118**, 10-24. https://doi.org/10.1016/j.compedu.2017.11.005

[26] Hone, K.S. and El Said, G.R. (2016) Exploring the Factors Affecting MOOC Retention: A Survey Study. *Computers & Education*, **98**, 157-168. https://doi.org/10.1016/j.compedu.2016.03.016

[27] Hellerstein, J.M. (2008) Quantitative Data Cleaning for Large Databases. United Nations Economic Commission for Europe (UNECE).

[28] Aggarwal, C.C. (2015) Outlier Analysis. In: *Data Mining*, Springer, Cham, 237-263. https://doi.org/10.1007/978-3-319-14142-8_8

[29] Kriegel, H.P., Kröger, P. and Zimek, A. (2010) Outlier Detection Techniques. *Tutorial at KDD*, **10**, 71-76.

[30] Chawla, S. and Sun, P. (2006) Outlier Detection: Principles, Techniques and Applications. *Proceedings of the* 10*th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD*), Singapore, 9-12 April 2006..

[31] Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) Data Clustering: A Review. *ACM Computing Surveys*, **31**, 264-323. https://doi.org/10.1145/331499.331504

[32] Ji, J.C., *et al.* (2012) A Fuzzy K-Prototype Clustering Algorithm for Mixed Numeric and Categorical Data. *Knowledge-Based Systems*, **30**, 129-135. https://doi.org/10.1016/j.knosys.2012.01.006

[33] Madhuri, R., *et al.* (2013) Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms. *ICT and Critical Infrastructure*: *Proceedings of the* 48*th Annual Convention of Computer Society of India-Vol II*, Visakhapatnam, 13-15 December 2013, 137-144.

[34] Palani, K. (2020) Identifying At-Risk Students in Virtual Learning Environment using Clustering Techniques. National College of Ireland, Dublin.

[35] Lugo, M.J.F., von Lücken, C. and Espinoza E.R. (2016) Sequencing Educational Con-

tents Using Clustering and Ant Colony Algorithms. In: Uskov, V., Howlett, R. and Jain, L., Eds., *Smart Education and E-Learning*, Springer, Cham, 375-385. https://doi.org/10.1007/978-3-319-39690-3_33

[36] Vineetha, Y. and Kishore, K.K. (2020) The Taxonomy: Health Monitoring System Using Machine Learning Techniques.

[37] Zhou, C.Y. and Huang, L.J. (2010) The Improvement of Initial Point Selection Method for Fuzzy K-Prototype Clustering Algorithm. 2010 2*nd International Conference on Education Technology and Computer*, Shanghai, 22-24 June 2010, 549-552.

[38] Jackson, A. and Mentzer, N. (2017) Cluster Analysis in Engineering Education. *ASEE Annual Conference and Exposition*, Columbus, 25-28 June 2017, Paper ID#18317.

[39] Antonenko, P.D., Toy, S. and Niederhauser, D.S. (2012) Using Cluster Analysis for Data Mining in Educational Technology Research. *Educational Technology Research and Development*, **60**, 383-398. https://doi.org/10.1007/s11423-012-9235-8

[40] Huang, Z.X. (1997) Clustering Large Data Sets with Mixed Numeric and Categorical Values. *Proceedings of the* 1*st Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD*), Trondheim, 24-27 June 1997, 21-34.

[41] Cao, F.Y., Liang, J.Y. and Bai, L. (2009) A New Initialization Method for Categorical Data Clustering. *Expert Systems with Applications*, **36**, 10223-10228. https://doi.org/10.1016/j.eswa.2009.01.060

[42] Rousseeuw, P.J. (1987) Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

[43] Lin, T.C. (2004) The Role of Higher Education in Economic Development: An Empirical Study of Taiwan Case. *Journal of Asian Economics*, **15**, 355-371. https://doi.org/10.1016/j.asieco.2004.02.006

[44] Mishra, A. and Agarwal, A. (2019) How Does Economic Expansion React to Educational Expenditure, Financial Development, and Financial Integration? A Nonlinear Granger Causality and Quantile Regression Analysis in an Asian Perspective. *International Journal of Education Economics and Development*, **10**, 276-293. https://doi.org/10.1504/IJEED.2019.10021225