Original Article

Access this article online

Quick Response Code:

Website: www.jehp.net

DOI: 10.4103/jehp.jehp_1466_21

Online assessment in two consequent semesters during COVID-19 pandemic: K-means clustering using data mining approach

Farshid Abedi[1], Batool Eghbali[2], Narjes Akbari[3], Ehsan Sadr[4], Fatemeh Salmani[5]

[1]Department of Infectious Diseases, School of Medicine, Infectious Diseases Research Center, Birjand University of Medical Sciences, Birjand, Iran, [2]Department of Community Medicine, School of Medicine, Birjand University of Medical Sciences, Birjand, Iran, [3]Department of Oral and Maxillofacial Medicine, School of Dentistry, Infectious Diseases Research Center, Birjand University of Medical Sciences, Birjand, Iran, [4]e-Learning Center, Birjand University of Medical Science, Birjand, Iran, [5]Department of Epidemiology and Biostatistics, School of Health, Social Determinants of Health Research Center, Birjand University of Medical Sciences, Birjand, Iran

Address for correspondence:
Dr. Fatemeh Salmani, School of Health, Social Determinants of Health Research Center, Birjand University of Medical Sciences, Birjand, Iran.
E-mail: salmany_fatemeh@yahoo.com

Received: 02-10-2021
Accepted: 24-12-2021
Published: 28-09-2022

Abstract:

BACKGROUND: Education and assessment have changed during the COVID-19 pandemic so that online courses replaced face-to-face classes to observe the social distance. The quality of online assessments conducted during the pandemic is an important subject to be addressed. In this study, the quality of online assessments held in two consecutive semesters was investigated.

MATERIALS AND METHODS: One thousand two hundred and sixty-nine multiple-choice online examinations held in the first (n = 535) and second (n = 734) semesters in Birjand University of Medical Sciences during 2020–2021 were examined. Mean, standard deviation, number of questions, skewness, kurtosis, difficulty, and discrimination index of tests were calculated. Data mining was applied using the k-means clustering approach to classify the tests.

RESULTS: The mean percentage of answers to all tests was 69.97 ± 19.16, and the number of questions was 34.48 ± 18.75. In two semesters, there was no significant difference between the difficulty of examinations (P = 0.84). However, there was a significant difference in the discrimination index, skewness, and kurtosis of tests (P < 0.001). Moreover, according to the results of the clustering analysis in the first semester, 43% of the tests were very hard, 16% hard, and 7% moderate. In the second semester, 43% were hard, 16% moderate, and 41% easy.

CONCLUSION: To evaluate the tests' quality, calculating difficulty and discrimination indices is not sufficient; many factors can affect the quality of tests. Furthermore, the experience of the first semester had changed characteristics of the second-semester examinations. To enhance the quality of online tests, establishing appropriate rules to hold the examinations and using questions with higher taxonomy are recommended.

Keywords:

Assessment, cluster analysis, COVID-19, discrimination, distance educationAci volorestet, arit

Introduction

Coronavirus disease (COVID-19) has affected many people worldwide so that 200 million people have been infected. One year after the outbreak of the disease, life has changed extensively in business, communication, education, and research dimensions. COVID-19 has had a profound effect on medical education, as some studies have suggested. Perhaps, the most important effect of this pandemic is the postponement of practical classes and internships. The length of the semester model, followed by the length of the students' studies, has been a challenge in itself.[1] Hence, this pandemic has developed virtualization in medical education.[2] Face-to-face courses and subsequent examinations were disrupted, followed by maintaining social distance and observing health protocols. Hence, universities replaced face-to-face

How to cite this article: Abedi F, Eghbali B, Akbari N, Sadr E, Salmani F. Online assessment in two consequent semesters during COVID-19 pandemic: K-means clustering using data mining approach. J Edu Health Promot 2022;11:307.

examinations with online tests.[3] Although online education and evaluation is not a new phenomenon, the recent epidemic outbreak has drawn scholars' attention toward it.[4] For example, Jagadeesan and Neelakanta (2021) are used an online self-assessment tool for medical students in biochemistry during pandemic.[5] Online tests, which are administered via the Internet over a period of time, are a reasonably effective approach to evaluate applicants' knowledge. In most cases, students must assemble in one location at the institution to take the examination; however, in online tests, all participants must be linked to the Internet and join the relevant website. However, sometimes, validity of these tests may be compromised due to the uncontrollable nature of test takers, which is one of the most important challenges in this type of test.[6]

The advantages of online examinations include easy access, time and cost savings,[7] instant posttest feedback, use of multimedia in designing questions, display of options according to personal preferences, increased creativity and thinking power, as well as clarity in receiving answers. Regarding multiple-choice questions having the highest rate of students' acceptance in online tests, these tests are suitable for assessing knowledge.[8] However, the disadvantages of online examinations are nonacceptance of technology, infrastructure problems, and the increased likelihood of cheating.[6]

To evaluate the effectiveness of training and assessment, it is necessary to analyze and evaluate the examination quality after its administration so that the organizers and participants can ensure its quality. In this regard, Salas-Morera *et al*. noted that online quizzes were effective in increasing students' academic performance.[9] Hingorjo and Jaleel and Mahjabeen *et al*. analyzed multiple-choice question tests based on the difficulty index, discrimination index, and distractor efficiency.[10,11] Upadhyah *et al*. examined multiple-choice tests by calculating differentiation index and discrimination index followed by drawing the scatter plot between them.[12]

Although difficulty and discrimination indices are often used to assess tests,[13] employing two indices to measure the quality of a test or its questions is insufficient. In this case, a collection of indicators is often employed to offer a holistic perspective of the test. We intend to describe the quality of tests held in consecutive semesters in the corona period using the indicators of the percentage of the mean score, standard deviation, number of questions, skewness, and kurtosis of answers. Furthermore, using k-means clustering method, we present the most important variables affecting the tests to classify the tests.

Materials and Methods

Study design

The present cross-sectional study was conducted on the examinations held during two consecutive semesters in the academic year 2020–2021 in Birjand University of Medical Sciences, including nine faculties (medicine, dentistry, pharmacy, health, paramedical, nursing, and midwifery in Birjand and nursing in Tabas, nursing and midwifery in Ghaen, and paramedical in Ferdows).

Sampling and data collection tool and technique

One thousand two hundred and sixty-nine online multiple-choice questions tests (535 tests related to the first semester and 734 tests related to the second semester) were included in the study. Quantitative variables, including the percentage of mean score, standard deviation, and number of questions, skewness, and kurtosis of answers, discrimination coefficient, and difficulty coefficient were calculated for each test, and the checklist of each test was completed.

Cheang and Hasni (1998) defined the difficulty index as the ratio of students who answered a question correctly to the total number of students taking the test. The discrimination index is also considered as the power of a test item; that is, the degree to which success or failure on an item indicates possession of the ability being measured. Moreover, skewness and kurtosis values were defined as deviation from normal distribution.[14]

The marginal mean of each test was used to indicate the students' performance. The data mining method was also applied to determine the relationship among the characteristics of tests. Data mining includes tools and techniques for "extracting knowledge from a large repository of data," in which various techniques such as clustering are used.[15] Clustering is an analytical method for high information dimensions. That categorizes information in such a way that the points within a cluster are similar and different from members of other clusters.[16] In the k-means method, the cluster centroid is representative of the cluster so that the distance among all members of the cluster and its centroid is minimized. As mentioned below:

$$J_K = \sum_{k=1}^{K} \sum_{i \in C_k} \left(x_i - m_k \right)^2$$

Where, $(x_1, x_{2,\ldots}\, x_n) = X$ is the data matrix and $m_k = \sum_{i \in C_k} \frac{x_i}{n_k}$ is the centroid of cluster C_k.

In k-means clustering algorithm, the following steps are performed: (1) K of the initial point is selected as the clusters' centroid, (2) all points are allocated to the nearest center of the cluster; (3) centers of the new

clusters are calculated; and (4) steps 2 and 3 are repeated until centroid of the clusters does not change.[17]

The rattle package was used to analyze information, identify clusters, and determine the relationships. Moreover, SPSS 19 (SPSS Inc., Chicago, Illinois, USA) and rattle package in R3.6.3 were used for data analysis. The significance level was considered at 0.05 in all tests.

Ethical consideration:

For ethical reasons, the tests were analyzed collectively.

Results

One thousand two hundred and sixty-nine tests conducted in the first (n = 535) and second (n = 734) semesters in Birjand University of Medical Sciences during 2020–2021were investigated. The mean percentage of correct answers to the questions of each test was 67.97 ± 19.16 and the mean number of questions was 34.48 ± 18.75. Negative skewness and kurtosis were reported for the tests, which were significantly different in two consecutive semesters (P < 0.001). The mean difficulty index was 0.67 ± 0.19 and the discrimination index was 0.32 ± 0.18. Table 1 entails details of the quantitative indices of the tests [Table 1].

Since the optimal difficulty index for multiple-choice questions tests is usually 0.625, it can be said that the studied examinations were at the desired level of difficulty. However, the observed low level of discrimination index was expected since in criterion-referenced tests, majority of students are able to answer most questions correctly.

The relationship between the difficulty and discrimination indices can indicate the quality of tests [Figure 1]. The inside area of the triangle illustrates the appropriate distribution of the tests so that 23% (295 tests) of all tests were out of the desirable area, which was 23% (123 tests) in the first semester and 23.7% (172 tests) in the second semester. However, the difference was not significant between the two semesters (P = 0.77).

Followed by describing the online tests, the relationship between test indices was examined. Since skewness and kurtosis of the tests were different, each semester was analyzed separately.

Based on the data mining analyses, which indicate the relationships between test indices [Figure 2], a positive and significant linear relationship was observed between skewness and kurtosis in the first semester [blue circles in Figure 2a]. However, a weak negative linear relationship was found among the number of questions, percentage of mean scores, and mean question difficulty index [Figure 2a. Pink circles]. In the second semester, the most important correlating variables included the number and difficulty of questions [Figure 2b]. Accordingly, the correlation pattern between test features has changed over time.

The studied tests were clustered using k-means clustering. Table 2 shows the characteristics of the test clusters in each semester. Tests of the first semester were divided into four clusters, in which 73.31% of the changes were defined using the percentage of mean scores and standard deviation of the tests.

Each cluster was named based on the values of its centroid, especially test difficulty and discrimination, as the two main components. Moreover, considering the number of tests in each cluster, 43%, 16%, 7%, and 34% of the tests were very difficult, difficult, moderate, and easy, respectively.

In the second semester, 100% change was observed only by considering the number of questions and test difficulty, which is consistent with the correlation matrix

Table 1: Characteristics of online tests and comparison of the results between two semesters

Variables	Total	Semester	Mean±SD	Test statistics	P
Average Percent	67.97±19.16	First	67.40±19.23	−0.91	0.37
		Second	68.39±19.12		
SD	3.14±1.97	First	3.18±1.98	0.61	0.54
		Second	3.11±1.97		
Number of question	34.48±18.75	First	35.50±18.75	1.65	0.10
		Second	33.74±18.74		
Skewness index	−2.64±14.07	First	−0.82±0.90	4.00	<0.001
		Second	−2.57±7.96		
Kurtosis index	−3.22±19.42	First	−0.990±1.15	3.55	<0.001
		Second	−2.98±9.11		
Difficulty index	0.67±0.19	First	0.67±0.19	−0.20	0.84
		Second	0.68±0.20		
Discrimination index	0.32±0.18	First	0.30±0.17	−2.99	<0.001
		Second	0.33±0.18		

SD=Standard deviation

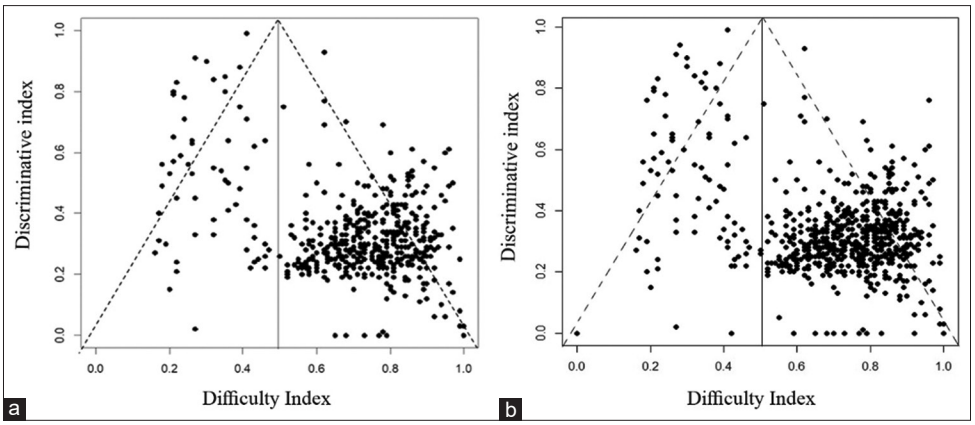

Figure 1: Scatter plot of the difficulty and discrimination indices in two consecutive semesters; (a) first semester, (b) second semester

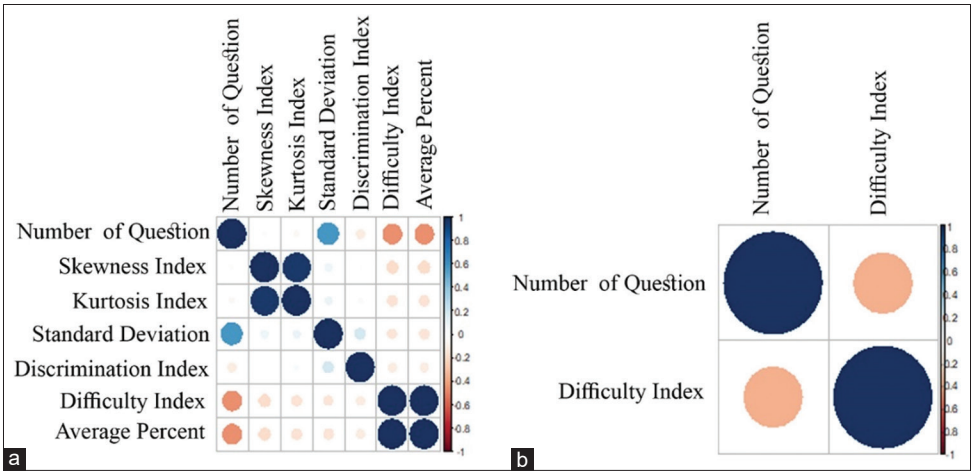

Figure 2: Correlation between characteristics of examinations in two consecutive semesters; (a) first semester, (b) second semester

Table 2: Clusters' centroid and goodness of fit in two consecutive semesters

Label of tests in each cluster	First semester				Second Semester		
	Cluster 1 Very difficult	Cluster 2 Difficult	Cluster 3 Moderate	Cluster 4 Easy	Cluster 1 Difficult	Cluster 2 Moderate	Cluster 3 Easy
Component for clustering							
Average percent	0.13	0.25	0.56	0.79	-	-	-
SD	0.23	0.16	0.29	0.16	-	-	-
Number of question	0.38	0.43	0.26	0.19	0.30	0.19	0.15
Difficulty index	0.14	0.25	0.57	0.79	0.29	0.65	0.84
Skewness index	0.81	0.69	0.82	0.75	-	-	-
Kurtosis index	0.74	0.60	0.74	0.67	-	-	-
Discrimination index	0.59	0.01	0.29	0.3	-	-	-
Percentage of online examinations in cluster	43	16	7	34	43	16	41
Sum of squares in cluster	7.26	9.01	1.87	7.12	1.89	3.41	2.98

Goodness of fit for first two components, First semester=73.31% and Second semester=100%. SD=Standard deviation

presented in Figure 2. According to the findings, 43% of the tests were difficult, 16% were moderate, and 41% were easy [Table 2].

The sum of squares in clusters of the second semester is less than the first semester and has better goodness of fit. This can be due to the presence of some other influential variables such as test time, infrastructure quality, and other hidden factors. Followed by changes in the pattern of tests in the second semester, only difficulty and number of tests were determined as effective factors in clustering the questions. This change can be caused by the students' experience from the first semester. Figure 3 illustrates the tests' distribution in clusters.

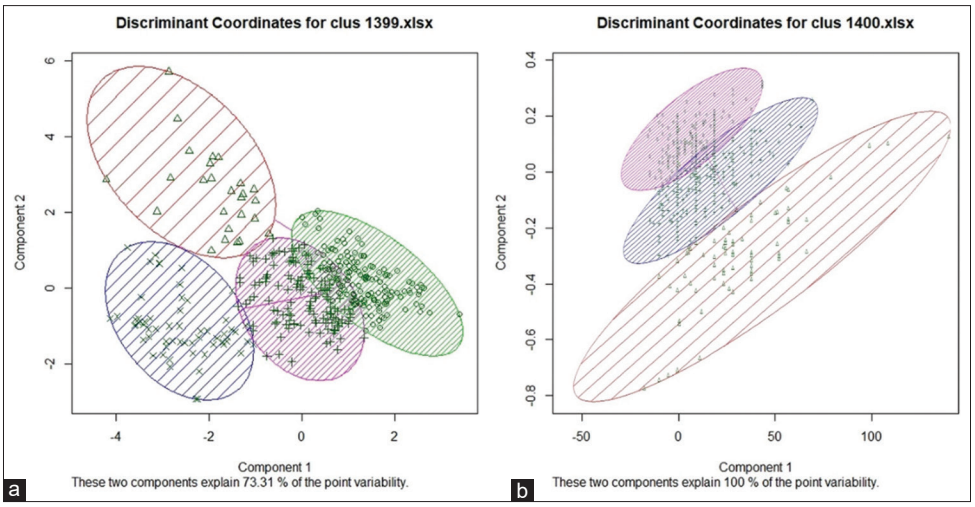

Figure 3: K-mean clusters for two consecutive semester examinations; (a) first semester, (b) second semester

Discussion

Assessment is making inferences about a student's learning outcome, which is usually in the form of "assessment of learning" (summative) and "assessment for learning" (formative). Summative method is usually used for final evaluation (pass/fail). Recently, assessment of the students' knowledge and skills was challenged due to the COVID-19 pandemic to maintain social distance.[18] Social distance implies that holding traditional tests with large crowds in test halls is impossible. Months after the onset of COVID-19 pandemic, students were repeatedly evaluated using a variety of methods, such as online tests. Therefore, quality assessment and analysis of the conducted tests are of particular importance due to the uncertainty in the persistence of the pandemic. These evaluations help decision-makers in education and assessment to conduct examinations with greater reliability and quality. In the present study, we described the tests in general and examined the trend of changes in test indices during two consecutive semesters. The difficulty index of questions was at an acceptable level, but the discrimination index of the tests was low.

Hassan *et al*. assessed the quality of online assessment among medical students during COVID19. They concluded that the mean scores of discrimination, difficulty, and student's scores increased significantly among the online multiple-choice questions tests.[19] These results were consistent with the present study discrimination index and inconsistent in the difficulty index report. Baghaei *et al*. evaluated the end-of-semester multiple-choice questions of nursing students and reported that the discrimination index was at the moderate level,[13] which is not consistent with our findings.

The mean percentage of correct answer tests was about 67%, which means that students who took the test were able to answer more than half of the questions on average. In other words, they were often able to pass the test. On the contrary, Holbrook *et al*. conducted a multi-year cross-sectional study, evaluated the online prescribing competence assessment among the final year Canadian medical students, and reported that the overall pass rate was 47.6%.[20] One explanation for the disparity in the proportion of mean scores is that in knowledge-based assessments, there is often a left skewness in the results, with students with higher scores outnumbering those with lower scores. In skill-based assessments, however, this occurrence is less typical. As a result, the passing rate in the cited study's skill test was lower than in ours. The scatter plot between difficulty and discrimination indices showed a nonlinear relationship, which can justify the lack of a strong linear relationship among these indices in the study. Similarly, Upadhyah *et al*. presented the nonlinear relationship between these two indices by drawing a scatter plot.[12] This difference may be due to the test format.

Since educators and students were less experienced with regard to the online examinations in the first semester, many factors affected the classification and quality of the tests. As one of the most important components in classifying the quality of tests, the level of difficulty also played a significant role in the second semester. This finding is consistent with the study by Johari *et al*., who reported that the difficulty index was directly related to the achievement of program outcomes.[21] In our study, in both semesters, the difficulty coefficient of the tests was higher than the Johari *et al*. study. This could be due to a change in test conditions. In the presence of a corona pandemic, the lack of control and security of the tests can increase the difficulty factor in our study.

The analysis of test characteristics over time was considered, which was lacked in the literature. Although

no significant change was observed in the students' mean scores over time, the kurtosis and skewness values were higher in the second semester than in the first semester. In other words, increase of the students' scores in the second semester had simplified the structure of relationship among test indicators in the second semester. This shows that the persistence of the pandemic has changed the experience of teachers, students, and officials, resulting in an alteration in the quality of examinations. These changes can be either due to teachers' improved skill in designing questions, officials' more careful planning in conducting examinations, or students' adaptation to the online training and examinations. The results of this study are important because it not only examines the changes in test indices during two consecutive semesters but also provides a more accurate classification of tests using statistical methods, while this is the case in similar articles, less seen.

Limitation and recommendation

The following are some of the limitations of the current research. We were unable to evaluate the influence of other variables on the participants' test performance since no information about their demographic characteristics was available throughout the testing. Regarding the changes in regulations of the virtual and timed examinations, educators and authorities are suggested to evaluate of next semester tests to assessment of the impact of new rules on holding examinations. It is also suggested that in addition to considering the demographic characteristics of students and teachers, the effect of the number of online, face-to-face, and offline classes was considered on the success of students.

Conclusion

To classify the tests, not only difficulty and discrimination properties but also other characteristics of tests should be investigated. Considering the impact of students' previous experience on their results in the second semester, using multi-purpose tools is recommended during the academic courses. In order to eliminate this impact, university instructors can consider various testing methods such as presenting conceptual or short-answer questions as well as randomizing questions and test options.

Acknowledgments:

The authors thank all people who helped us gather the information. This plan with the code IR.BUMS.REC.1399.437 has been approved by the Ethics Committee of Birjand University of Medical Sciences.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Rezaei H, Haghdoost A, Javar HA, Dehnavieh R, Aramesh S, Dehgani N, *et al.* The effect of coronavirus (COVID-19) pandemic on medical sciences education in Iran. J Educ Health Promot 2021;10:136.
2. Ghadrdoost B, Sadeghipour P, Amin A, Bakhshandeh H, Noohi F, Maleki M, *et al.* Validity and reliability of a virtual education satisfaction questionnaire from the perspective of cardiology residents during the COVID-19 pandemic. J Educ Health Promot 2021;10:291.
3. Harries AJ, Lee C, Jones L, Rodriguez RM, Davis JA, Boysen-Osborn M, *et al.* Effects of the COVID-19 pandemic on medical students: A multicenter quantitative study. BMC Med Educ 2021;21:14.
4. Khan RA, Jawaid M. Technology Enhanced Assessment (TEA) in COVID 19 pandemic. Pak J Med Sci 2020;36:S108-10.
5. Jagadeesan AR, Neelakanta RR. Online self-assessment tool in Biochemistry – A medical student's perception during COVID-19 pandemic. J Educ Health Promot 2021;10:137.
6. Walsh K. Point of view: Online assessment in medical education-current trends and future directions. Malawi Med J 2015;27:71-2.
7. Boitshwarelo B, Reedy AK, Billany T. Envisioning the use of online tests in assessing twenty-first century learning: A literature review. Res Pract Technol Enhanc Learn 2017;12:16.
8. Marius P, Marius M, Dan S, Emilian C, Dana G. Medical students' acceptance of online assessment systems. Acta Med Marisiensis 2016;62:30-2.
9. Salas-Morera L, Arauzo-Azofra A, García-Hernández L. Analysis of online quizzes as a teaching and assessment tool. J Technol Sci Educ 2012;2:39-45.
10. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. J Pak Med Assoc 2012;62:142-7.
11. Mahjabeen W, Alam S, Hassan U, Zafar T, Butt R, Konain S, *et al.* Difficulty index, discrimination index and distractor efficiency in multiple choice questions. Ann PIMS Shaheed Zulfiqar Ali Bhutto Med Univ 2017;13:310-5.
12. Upadhyah AA, Maheria PB, Patel J. Analysis of one best MCQS in five preuniversity physiology examinations. Int J Physiol 2019;7:10-5.
13. Baghaei R, Shams S, Feizi A. Evaluation of the nursing students final exam multiple-choice questions in urmia university of medical sciences. Nurs Midwifery J 2016;14:291-9.
14. Cheang Q, Hasni Z. Analisis item dan pembinaan ujian-satu perbandingan antara pendekatan rujukan norma dan rujukan kriteria. J Pendidikan Tigaenf 1998;2:112-20.
15. Verma M, Mehta D. A comparative study of techniques in data mining. Int J Emerg Technol Adv Eng 2014;4:314-21.
16. Dubey A, Choubey A. A systematic review on k-means clustering techniques. Int J Sci Res Eng Technol 2017;6:624-7.
17. Adriyendi M. Clustering using K-means and fuzzy C-means on food productivity. International Journal of u- and e- Service, Science and Technology 2016;9:291-308.
18. Fuller R, Joynes V, Cooper J, Boursicot K, Roberts T. Could COVID-19 be our 'There is no alternative'(TINA) opportunity to enhance assessment? Med Teach 2020;42:781-6.
19. Hassan B, Shati AA, Alamri A, Patel A, Asseri AA, Abid M, *et al*. Online assessment for the final year medical students during COVID-19 pandemics; the exam quality and students' performance. Onkol Radioter 2020;14:1-6.

20. Holbrook A, Liu JT, Rieder M, Gibson M, Levine M, Foster G, *et al*. Prescribing competency assessment for Canadian medical students: A pilot evaluation. Can Med Educ J 2019;10:e103-10.

21. Johari J, Sahari J, Abd Wahab D, Abdullah S, Abdullah S, Omar MZ, *et al*. Difficulty index of examinations and their relation to the achievement of programme outcomes. Proc Soc Behav Sci 2011;18:71-80.