Strides Dev Med Educ. 2022 December; 19(1):150-154. doi 10.22062/sdme.2021.196760.1086

Brief Report

Assessing the Quality of Electronic Exams During the COVID-19 Pandemic

Zahra Zahedi[1], Hamid Salehiniya[2], Afagh Zarei[3], Hamid Abbaszadeh[4*]

[1]Student Research Committee, Birjand University of Medical Sciences, Birjand, Iran.
[2]Social Determinants of Health Research Center, Birjand University of Medical Sciences, Birjand, Iran.
[3]Department of Medical Education, Tehran University of Medical Sciences, Tehran, Iran.
[4]Department of Oral and Maxillofacial Pathology, Faculty of Dentistry, Birjand University of Medical Sciences, Birjand, Iran.

Received: 2021 October 20
Revised: 2021 November 23
Accepted: 2021 December 22
Published online: 2022 May 17

***Corresponding author:**
Department of Oral and Maxillofacial Pathology, Faculty of Dentistry, Birjand University of Medical Sciences, Birjand, Iran.
E-mail:
hamidabbaszade@yahoo.com

Citation:
Zahedi Z, Salehiniya H, Zarei A, Abbaszadeh H. Assessing the quality of electronic exams during the COVID-19 pandemic. Strides Dev Med Educ. 2022 December; 19(1): 150-154.
doi:10.22062/sdme.2021.196760.1086

Abstract

Background: Due to the widespread utilization of electronic exams, neglecting their quality is a major threat. Assessing the quality of electronic exams plays a decisive role in determining the efficacy of electronic learning.

Objectives: This study aimed to assess the quality of electronic exams held at the beginning of the coronavirus disease 2019 outbreak.

Methods: Following a cross-sectional design, this study included all electronic exams of the electronic test center of Birjand University of Medical Sciences during the academic year of 2020. Reliability, discrimination index (DI), and difficulty index (DIF) of exams were used to assess the quality. Descriptive statistics and frequency distributions were used to describe the data.

Results: Out of 101 E-exams, 59.4% had appropriate DIF, 61.4% had low DI, and 66.3% had unfavorable reliability. Also, 38.6% of exams had high DIF (easy questions). For all exams, the mean of DIF, DI, and reliability was 0.66±0.14, 0.28±0.08, and 0.56±0.31, respectively. The mean of DI (P= 0.30) and reliability (P= 0.09) was not significantly different based on faculty. The mean of DIF was significantly different according to the faculty (P=0.03).

Conclusion: Concerning the quality of e-exams, most problems are related to the DI and reliability. It is recommended to hold empowerment workshops on how to design exam questions for faculty members to get them acquainted with strategies to increase the reliability and discrimination index of the exam.

Keywords: Assessment, Electronic, Quality, Evaluation, Analysis

Background

Utilization of educational technologies by universities has significantly increased in the last few years due to its potential to enhance learning and teaching outcomes, which in turn results in several benefits for teaching and learning. Electronic exams (E-exams) are computer-based exams. E-exams are considered a major transformation for education in universities (1). E-exam has several advantages, including the easiness to use, getting instant results, ability to provide answers at the end, better interpretation and analysis of the results, multiple capabilities in using text, images, audio, and video, save on paper, and ability to improve the assessment quality. Electronic tests provide the possibility of cooperation between universities (2-7).

Assessing the quality of electronic exams plays a decisive role in determining the efficacy of electronic learning, which has attracted insufficient attention. Regrading increasing inclination toward e-exams, neglecting their quality is a major threat (8). There are different definitions for quality. Some have mentioned quality as a subjective issue that can be elicited from the audience's point of view and their level of satisfaction. On the other hand, some believe that quality is associated with objectivity and have used quantitative criteria and relevant standards to assess quality. Reliability, discrimination index (DI), and difficulty index (DIF) are among the quantitative criteria developed for evaluating the quality of electronic tests (8-10). For example, some studies have evaluated the quality of e-exams by subjective methods (based on students' questionnaires) (6, 8). Several studies have used objective item analysis (e.g., DIF and DI items) to evaluate the quality of paper-based exams (11-14). Only a few studies have evaluated the quality of e-exams based on objective items (DIF and DI items) (10). Noteworthy, these few studies have mostly focused on the quality of online e-exams, and insufficient attention has been paid to the evaluation of the quality of isolated (campus-based) e-exams.

Therefore, due to the widespread use of electronic exams, this study aimed to assess the quality of

electronic exams held during the COVID-19 pandemic at the electronic exam center of Birjand University of Medical Sciences (BUMS).

Objectives

This study aimed to assess the quality of electronic exams held at the beginning of the coronavirus disease 2019 outbreak.

Methods

Following a cross-sectional design, this study included all e-exams of the electronic test center of BUMS. The study was confirmed by the ethical committee (ethical number: IR.BUMS.REC.1399.076). The sampling method was census. E-exams with multiple-choice questions (MCQ) were included in this study. The exclusion criteria were paper-based exams, exams held in other universities, exams held outside the electronic exam center of BUMS, exams held on academic years other than 2019-2020, and other types of exams except for MCQ.

To evaluate the quality of exams, we assessed the mean DIF, the mean DI, and reliability of tests performed during the academic year of 2019-2020. All of this information was extracted from the database of the Electronic Exam Center of BUMS.

The DI determines the strength of the item in distinguishing between the strong group and the weak group of students, which is a number between -1 to +1. The higher this index, the more desirable it is. The analysis of the descriptive discrimination index in the Electronic Exam Center system of the university was such that discrimination index greater than 0.3 was considered "appropriate" and less than 0.3 as "low". Obviously, the closer the value of this index to +1, the more powerful the test items are for distinguishing between strong and weak students (9).

The DIF indicates the percentage of correct answers, which ranges from zero to +1. A difficulty index of 0.3-0.7 indicates the appropriateness of the item or exam. Meanwhile, a value less than 0.3 indicates "difficulty", and a value higher than 0.7 shows "easiness" of the item or exam (9).

Reliability refers to the accuracy, stability, or repeatability of test results, which is usually determined by Cronbach's alpha; the closer the number to one, the greater the internal correlation among items, indicating higher homogeneity of the items of a test (9). The reliability of the tests in the University Electronic Exam Center system was calculated by calculating Cronbach's alpha. Cronbach suggested a reliability coefficient of 0.45 as "low", 0.75 as "average, and acceptable" and 0.95 as "high" (9). According to the University Electronic Exam Center system, values higher than 0.7 were considered favorable.

Due to the low number of exams held in the two faculties of Nursing and Midwifery and Paramedical, we merged the exams of these two faculties. The collected and analysis data were administered by SPSS version 16, descriptive statistics and frequency distributions, for example the mean and standard deviation were used to describe the data.

Results

In this study, 101 e-exams held in the Electronic Exams Center of the university in the academic year of 2019-2020 were reviewed. The highest number of e-exams was related to the medical school (n=71; 70.3%), followed by the dentistry school (n=19; 18.81%). On the other hand, the lowest number of exams was related to the paramedical school (n=1; 0.99%) and nursing and midwifery schools (n=2; 1.98%). Eight exams (7.92%) were held in the Faculty of Public Health.

For all exams, the mean value of DI, DIF, and reliability of the exams was, respectively, 0.28±0.08, 0.66±0.14, and 0.56±0.31. Table 1 shows the mean value of DIs, DIF and reliability of exams held by schools of the BUMS. The mean value of DI of the exams held by Schools of Nursing and Midwifery, Paramedical, and Public Health was appropriate. While schools of Medicine and Dentistry obtained 'inappropriate' values for this index. The highest value of DI was related to exams held by the School of Public Health, and the lowest value belonged to the School of Dentistry.

Table 1. The mean discrimination indexes, difficulty indexes, and reliability of exams held by different faculties

Item / School	Discrimination index		Difficulty index		Reliability	
	Mean (SD)	Description	Mean (SD)	Description	Mean (SD)	Description
Medical	0.28 (0.88)	Low	0.64 (0.15)	Appropriate	0.59 (0.24)	Low
Dentistry	0.26 (0.05)	Low	0.66 (0.09)	Appropriate	0.54 (0.27)	Low
Health	0.32 (0.09)	Appropriate	0.79 (0.13)	Easy	0.31 (0.69)	Low
Nursing and Midwifery+ Paramedical	0.30 (0.08)	Appropriate	0.66 (0.09)	Appropriate	0.76 (0.28)	Appropriate

The mean DIF of the exams was appropriate in all schools, except for the School of Public Health, which the items were considered 'easy ' according to the DIF. The highest DIF of exams (easiness of items and

exams) was related to the School of Public Health and, the lowest value (i.e., higher difficulty of items and exams) was related to the School of Nursing and Midwifery and School of Paramedical.

The mean reliability of the exams was favorable for the School of Paramedical and School of Nursing and Midwifery. On the other hand, it was unfavorable for other schools. The highest mean reliability of exams was related to the School of Paramedical and School of Nursing and Midwifery. Meanwhile, the lowest mean reliability was related to the School of Public Health.

DI of 39 exams (38.6%) was "appropriate", and the mean DI of 62 exams (61.4%) was "low". For 60 exams (59.4%), the DI was "appropriate", "easy" for 39 exams (38.6%), and "difficult" for two exams (2%). The reliability of 34 exams (33.7%) was "favorable", and the reliability of 67 exams (66.3%) was "unfavorable".

Table 2 presents the situation of exams held by various schools, separated by different categories of DI, DIF, and reliability.

Table 2. Frequency of exams of schools in different categories of discrimination index, difficulty index, and reliability

Item / School	Number of exams N (%)	Discrimination index		Difficulty index			Reliability	
		Appropriate N (%)	Low N (%)	Easy N (%)	Appropriate N (%)	Difficult N (%)	Favorable N (%)	Unfavorable N (%)
Medical	71 (70.3)	28 (39.4)	43 (60.6)	27 (38)	42 (59.2)	2 (2.8)	27 (38)	44 (62)
Dentistry	19 (18.81)	5 (26.3)	14 (73.7)	6 (31.6)	13 (68.4)	0 (0)	4 (21.1)	15 (78.9)
Health	8 (7.92)	4 (50)	4 (50)	5 (62.5)	3 (37.5)	0(0)	1 (12.5)	7 (87.5)
Nursing and Midwifery + Paramedical	3 (2.97)	2 (66.7)	1 (33.3)	1 (33.3)	2 (66.7)	0 (0)	2 (66.7)	1 (33.3)

Discussion

This study aimed to assess the quality of electronic exams held at the beginning of the coronavirus disease 2019 outbreak. DIF, DI, and reliability were considered to assess the quality of e-exams. For e-exams that were held in the electronic exam center of BUMS, the mean DI of all exams was low. The mean DI of all exams was appropriate. The mean reliability of all exams was unfavorable. Abualrob et al. (2019), which intended to assess the quality of electronic tests at Arab American University Palestine (AAUP), reported insufficient assessment of exams' quality (6). In their study, students gave a moderate score to the quality of e-exams.

In the study by Pourafshar et al. (2020), DIF of face-to-face and online tests were, respectively, 0.62 and 0.68 (10). These findings are consistent with our finding, which DI of face-to-face and online tests was, respectively, 0.30 and 0.33. Although DI of their study is similar to our study; to some extent, there is no agreement between our findings and their findings because by definition, DI of their study is "appropriate", while in our study it was "low". This inconsistency can be attributed to the higher ability of university professors of the Kerman University of Medical Sciences in designing exam's items than their counterparts at BUMS. They concluded that since face-to-face and online tests were considered appropriate, based on DI and DIF criteria, it appears that e-tests may be an appropriate alternative for face-to-face tests.

In a study by Musa et al. (2018) on physiology multiple choice question (MCQ) tests at Khartoum University, the mean DIF index was 0.56 and the majority of items had acceptable difficulty (11). In this respect, their findings are in line with this study. With respect to DI, 90.1% of items were acceptable.

Also, there is a discrepancy between the findings of the present study and their study, which can be attributed to designing strongly difficult/easy items by university professors of BUMS, which led to a relatively low DI of the exams. Ganji Arjenaki (2017) reported a positive and significant association between students' satisfaction and the quality of e-exams (8).

To assess the quality of e-exams, they evaluated the quality of the evaluation criteria, the quality of counter-fraud, the quality of learning, the quality of using new learning methods, and the quality of providing information on all aspects of the test. There was a positive and significant association between all items (except for the quality of counter-fraud) and student satisfaction. Noteworthy, in comparison to our study, they used different criteria to assess the quality of e-exams; hence, the findings of the two studies are not comparable.

In a study by Taib and Yusoff (2014) on MCQ in paper-based exams of fourth-year medical students, MCQ's DIFs ranged from 0.67 to 0.79 (the level was appropriate). This finding is in line with our results, except that they investigated paper exams. MCQs showed high DI (0.58-0.76), which implied its higher appropriateness for discrimination among students

(12). In the present study, the DI was low and ranged from 0.22 to 0.35. The results of our study do not match their results, which can be attributed to the higher ability of teachers to design standard items in their study. On the other hand, the nature of the two studies is different, i.e., we investigated e-exams.

Boopathiraj and Chellamani (2013) performed a study to assess items of an exam in the education field and showed that most of the items had proper DIF and DI. However, some items were not accepted because of inappropriate DI (13). Our results are in line with their results in the sense that in both studies, the weakness is mainly observed in the DI of items rather than the DI.

In a study by Mahjabeen et al. (2017), according to DIF, out of total 65 exams, 81% of MCQs were acceptable, 2% low DIF, and 17% had high DIF (14).

According to DI, 62% had very good DI; 23%, 8%, and 17% had good, acceptable, and poor DI, respectively. In our study, 59.4% of e-exams had appropriate DIF, which due to the fact that they had a higher percentage, compared to easy and difficult items, is in accordance with their result. In our study, 61.4% of e-exams had low DI, which contradicts the findings of their study and shows a significant difference with their result. The cause of this difference is the ability of teachers to design standard items in the two studies; that is, the higher ability of their teachers in designing items. Of course, it should not be overlooked that the nature of the two studies is somewhat different because we investigated electronic tests and they analyzed paper tests.

It is necessary to mention some limitations and biases of our study, including not comparing the quality of paper-based and e-exams. Hence, further studies should compare the quality of paper-based and e-exams using a similar sample of students. Another limitation is not comparing the quality of on-campus (isolated) e-exams and home-based (non-isolated) e-exams, which is suggested to include in future studies due to the extensive use of online exams. Another limitation of this study is the small number of e-exams held in some schools; hence, the authors recommend performing future studies on larger sample size (more e-exams).

It is suggested to hold faculty development programs for faculty members to improve their skills on how to design standard items with appropriate discrimination index and reliability.

Conclusion

In general, major problems in the quality of electronic exams are associated with DI and reliability of exams; the DIF is generally appropriate. E-exam should be monitored continuously, and feedback should be provided to faculty members.

Acknowledgements: We thank the Research Vice-Chancellor of BUMS for supporting this research.

Conflict of interests: We have no conflict of interest.

Ethical approval

This study has been approved by the ethical committee of BUMS (code: IR.BUMS.REC.1399.076).

Funding/Support: There was no financial support.

References

1. Wibowo S, Grandhi S, Chugh R, Sawir E. A pilot study of an electronic exam system at an Australian University. Journal of Educational Technology Systems. 2016; 45(1):5-33. doi:10.1177/0047239516646746.
2. Llamas-Nistal M, Fernández-Iglesias MJ, González-Tato J, Mikic-Fonte FA. Blended e-assessment: migrating classical exams to the digital world. Computers & Education. 2013; 62(1):72-87. doi:10.1016/j.compedu.2012.10.021.
3. Merdzhanov I. Advantages of the electronic exam. Knowledge International Journal. 2019; 35(2):553-8.
4. Amer ME. Effectiveness of using electronic exams in assessment in Saudi universities: empirical study. International Journal of Educational Technology and Learning. 2020 Jun 25; 8(2):61-9. doi:10.20448/2003.82.61.69.
5. Torssonen T. Electronic exam offers opportunities for collaboration and flexibility in Finnish higher education institutes. Proceedings of the EdMedia+ Innovate Learning conference; 2020 June 1-5; Amsterdam, Netherlands.
6. Abualrob MMA, Asad NAA, Abu Daqar MAM. Attitudes toward and implications of the computer-based exams at Arab American University of Palestine. Journal of Education and Learning. 2019; 8(1):196-205. doi:10.5539/jel.v8n1p196.
7. Biantoro B, Arfianti A. Issues in the Implementation of computer-based national exam (CBNE) in Indonesian secondary schools. Proceedings of the Third International Conference on Sustainable Innovation 2019–Humanity, education and social sciences (IcoSIHESS 2019); 2019 July 30-31; Yogyakarta, Indonesia. doi: 10.2991/icosihess-19.2019.69. [PMID:31406532]. [PMCID:PMC6684423]
8. Ganji Arjenaki B. Surveying the quality of electronic tests in the student satisfaction. Med Sci. 2017; 10 (3): 180-8. [In Persian]
9. Seif A. Educational measurement, assessment and evaluation. 7th ed. Tehran: Doran Pub; 2019. [In Persian]
10. Malek Pourafshar R, Shojaeipour R, Khazaeli P, Bazrafshan A, Beigzadeh A, Dehghani MR. Comparison of analytic indices of in-person vs. online exams in an Iranian medical university in the academic year 2020. Strides Dev Med Educ. 2020 Sep 1; 17(Suppl): e91451.
11. Musa A, Shaheen S, Elmardi A, Ahmed A. Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine Khartoum University. Khartoum Medical Journal. 2018; 11(2): 1477-86.
12. Taib F, Yusoff MS. Difficulty index, discrimination index, sensitivity and specificity of long case and multiple-choice questions

to predict medical students' examination performance. Journal of Taibah University Medical Sciences. 2014; 9(2):110-4. doi:10.1016/j.jtumed.2013.12.002.
13. Boopathiraj C, Chellamani K. Analysis of test items on difficulty level and discrimination index in the test for research in education International Journal of Social Science & Interdisciplinary Research. 2013 Feb; 2(2):189-93.
14. 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(4):310-5.