Journal of Taibah University Medical Sciences (2022) 17(1), 128–136

Taibah University

Journal of Taibah University Medical Sciences

www.sciencedirect.com

Original Article

Web-based and paper-based examinations: Lessons learnt during the COVID-19 pandemic lockdown

Mohamed Nor-El-Din Saleh, PhD [a], Tarek Abdul Ra'oof Salem, PhD [b], Ahmad Saleh Alamro, PhD [c] and Majed Mohammed Wadi, MSc [c,*]

[a] *Anatomy Department, College of Medicine, Qassim University, KSA*
[b] *Pathology Department, College of Medicine, Qassim University, KSA*
[c] *Medical Education Department, College of Medicine, Qassim University, KSA*

Received 18 May 2021; revised 6 September 2021; accepted 12 September 2021; Available online 16 October 2021

الملخص

أهداف البحث: تصف هذه الورقة عملية تقييم الطلبة عن بعد في التعليم الطبي أثناء فترة الحظر في جائحة كوفيد-١٩، وتشارك الخبرات المستندة إلى البيانات في حل المشكلات الناشئة عن ذلك.

طرق البحث: قمنا بتحليل بيانات الاختبارات الورقية النهائية والاختبارات المستندة على الويب، التي أجريت على مدار العام الدراسي ٢٠١٩/٢٠٢٠. وتم تضمين اثني عشر اختبارا، أربعة اختبارات لكل مستوى دراسي، من السنة الأولى وحتى الثالثة. منها ثمانية اختبارات كانت ورقية، وأربعة اختبارات مستندة على الويب. قارنا متوسط درجات كل نوع من الاختبارات، وبين الاختبارات والمستوى الدراسي. بالإضافة إلى ذلك، قمنا بمقارنة درجات الاختبارين الورقي والمستند على الويب التي حصل عليها الطلاب العشرة الأوائل والعشرة الطلاب الأدنى تحصيلا.

النتائج: تم العثور على اختلافات في درجات الطلاب من كل دفعة من المجموعات الثلاث في الاختبارات المختلفة، سواء كانت ورقية أو مستندة على الويب. في بعض الحالات، كان الفرق ذا دلالة إحصائية. ولم يتم العثور على اتجاه / نمط محدد للاختلاف بين الدرجات في أي نوع من الاختبارات. كما كان متوسط الدرجات في الاختبارات المستندة على الويب وسطا بين المتوسطات الحسابية لطلاب السنة الأولى والثانية، ولكن أقل بالنسبة لطلاب السنة الثالثة. وأظهرت علامات الطلاب الفردية في الاختبارات المختلفة ارتباطا إيجابيا. وكان معامل الارتباط للاختبارات الورقية مرتفعا دائما.

الاستنتاجات: كشفت الدراسة الحالية عن عدم وجود فرق ملحوظ في نتائج الاختبارات الورقية والمستندة على الويب، سواء في متوسط الفصل أو لنتائج الطلاب الفرديين. على الرغم من وجود بعض الاختلافات بين نتائج نهجي التقييم، لم يكن هناك اتجاه ملحوظ.. ستوفر الاختبارات المستندة على الويب نهجا مثاليا للتقييم التكويني، والاختبار التحصيلي، والتقييم المتواصل.

الكلمات المفتاحية: التقييم؛ كوفيد-١٩؛ الاختبار الورقي؛ الاختبار المستند على الويب؛ الدرجات

Abstract

Objectives: This study describes the process of remote assessment in medical education during the COVID-19 lockdown and shares data-driven experiences in resolving emerging concerns.

Methods: We analysed the data of end-of-course paper-based exams (PBEs) and web-based exams (WBEs) conducted during the academic year 2019/2020. Twelve end-of-block exams were included. There were four exams each for the first-, second-, and third-year students. Eight exams were conducted as PBEs, and four were administered as WBEs. We compared the mean scores of PBEs and WBEs between exams and batches. Additionally, we compared the PBE and WBE scores obtained by 10 high-performance and 10 lowest-achieving students.

Results: Variations were found in the scores of students from each of the three batches in PBEs or WBEs. In a few instances, the difference was statistically significant. No specific trend or pattern was detected in the difference between the scores of PBEs and WBEs. The mean score for the WBEs was intermediate among the means of PBEs for the first- and second-year students, but lower for the third-year students. Individual students' marks in different exams consistently showed a positive

* Corresponding address: Medical Education Department, College of Medicine, Qassim University, KSA.
E-mail: m.wadi@qu.edu.sa (M.M. Wadi)

Peer review under responsibility of Taibah University.

ELSEVIER Production and hosting by Elsevier

 https://doi.org/10.1016/j.jtumed.2021.09.004

correlation. The correlation was always high for PBEs ($r = 0.782, 0.847$).

Conclusion: The present study showed that average and individual scores in WBEs and PBEs are comparable. Although there were some variations between the results of the two assessment modalities, no remarkable trend or pattern was observed. WBEs offer an ideal approach for formative assessment, progress testing, and the low-weight, but frequent, nature of continuous assessment.

Keywords: Assessment; COVID-19; Paper-based exam; Scores; Web-based exam

Introduction

The COVID-19 pandemic had a significant impact on all aspects of life, including the education sector. Commitment to social distancing led to the closure of almost all educational institutions worldwide to contain the spread of the disease.[1,2] However, medical education had to continue despite the threat,[3] necessitating a shift to online/distance learning.[4]

This sudden and rapid shift to distance learning was implemented by medical schools worldwide, which posed many challenges in medical education. The most salient obstacle was teaching in actual clinical settings,[5] which was tackled by restricting the number of students during bedside teaching.[6–8] Similarly, online assessment remains challenging.[9]

During e-assessment, many issues need to be considered, such as, selecting the most appropriate assessment modality; adapting a feasible and 'user-friendly' platform acceptable to students and teachers; and maintaining validity by maximising assessment security and minimising malpractice.[9,10]

Dennick et al.[11] described numerous practical tips and provided a comprehensive foundation to establish an e-assessment centre. Their contributions are extremely useful for e-assessment in well-equipped computer laboratories with strict measures to prevent cheating and test item leakage.

Understandably, e-assessment centres cannot be established in all medical colleges within due time. Even if this were possible, the use of computer laboratories would contravene mandatory social-distancing measures.[12] Therefore, local actions—with in-built flexibility—should be considered to develop the best possible practices for the e-assessment of students' achievements.[13]

The Ministry of Education (MOE) in KSA announced e-assessment regulations for all universities in the Kingdom for the second semester of the academic year 2019/2020. It mandated 80% of the assessment weightage to be assigned to continuous assessment, and the remaining 20% to the final test targeted at the cognitive domain.[14] This approach is scientifically sound; studies reveal that numerous formative assessment tasks and frequent quizzes engage students for better learning.[6,9] Moreover, assessment of the cognitive domain is possible through e-assessment modalities including multiple-choice questions (MCQs), modified essay questions, assignments, and open-book exams. These modalities can be implemented using a variety of learning management system (LMS) platforms, such as Blackboard and Moodle.[10]

This study reflects Qassim University College of Medicine (QUCOM)'s experience in conducting remote written e-assessments during the COVID-19 lockdown, specifically highlighting the concern related to the validity of web-based exams (WBEs) in terms of similarity of results to that of paper-based exams (PBEs), and issues of test security.

Materials and Methods

The present study utilised data from written end-of-course exams in the preclinical phase (first, second, and third years) composed of integrated transdisciplinary body system blocks, in the QUCOM during the academic year 2019/2020. Students of each level took four exams during the academic year. Eight of the 12 exams were PBEs, and were conducted before the COVID-19 lockdown; they comprised an MCQ component and a constructed-response component (i.e., modified-essay questions and short-essay questions). The MCQ component lasted 2 h, comprised 100 type-A MCQs (containing a vignette, a lead-in-question, and four options), and represented 50% of total marks.

The four remaining exams were conducted online during the lockdown in the MCQ format. Each WBE lasted 50 min and comprised 40 questions. In both situations, subject-matter experts in the college prepared the exam questions in accordance with the regulations of the institution's assessment unit. Each item comprised a short vignette, a lead—in statement, and four options: one correct choice and three distractors. The tests were subjected to internal moderation.

Measures to prevent cheating in the context of the current study included the following. First, assessment tasks were distributed throughout the entire period of distance learning with less weight for the end-of-course exam. Second, time-locked tests composed of type-A MCQs were used to ensure that answers were vignette-dependent and the lead—in alone provided no clue. Questions appeared at random to different students so that students could not join as a group to answer. Additionally, students were not allowed to return to a question after they answered it. Given the cultural constraints and Qassim University regulations preventing the viewing of girls, use of webcam to invigilate during the exam was not possible.

The QUCOM began with the digitalisation of learning well before the pandemic. Biweekly online formative assessments have been conducted since 2009 using the Moodle platform. In 2016, the college adopted the Blackboard LMS and added interactive sessions through distance learning, enjoying the infrastructure support provided by the university.

As a response to the instructions of the MOE and university regulations regarding shifting to distance learning,

the college formulated the E-assessment Committee (EAC) to conduct and manage remote assessment. The committee included representatives of different disciplines, moderators of the LMS in the college, and the responsible body for assessment (i.e., the Assessment Unit). Through online meetings, the EAC discussed all concerns and issues related to distance assessment to ensure feasibility, maximise validity, follow university regulations, and plan for the management of possible internet connection problems.

The committee decided that the final exam would be conducted as a one-part exam comprising 40 type-A MCQs with a duration of 50 min. Students would have to join the test within the first 15 min, could view only one question at a time, and would not be able to return to any question once an answer was submitted. For each batch, one mock test and two quizzes with 20 MCQs each were conducted with the same format and mode of implementation as the final exam.

The mock exam aimed to familiarise students and faculty with the process, disclose unforeseen obstacles, and provide clues to resolve any encountered challenges. The quizzes were intended to ensure the continuity of the learning process, assure students about their achievement, and minimise/alleviate their stress from high-stakes end-of-course exams during the pandemic.[15]

Students were instructed that all exams would be conducted through the official blackboard of Qassim University. Each student would access the exams using their own ID and password to log in. The university's tips and regulations for proceeding through exams were posted on the university' website[16] and sent via email to all students in Arabic—students' mother tongue—to avoid any misunderstanding that may hinder their performance during the e-assessment. These regulations included:

1. Preferably use your own computer to answer the questions.
2. If you want to use your smart phone, please use the available internet explorer, enter the university site, and select e-learning.
3. Please use Chrome explorer.
4. Do not enter the exam site before the announced time because it can refresh the exam start time on the website.
5. Do not press the 'start' button unless you have read all the exam instructions carefully.
6. You can see the answered as well as remaining questions by pressing the 'questions completion status' button.
7. Answer all questions within the allowed time as the exam will end at its set time.
8. Do not press the 'back' button from your internet explorer nor from the exam site as this will close the exam automatically.
9. Do not press the 'complete' button until you are sure that you have answered all questions.

Data from the end-of-course PBEs and WBEs conducted during the academic year 2019/2020 were analysed to explore the effectiveness of remote assessments. The similarity of the results between WBEs and PBEs was tested, and possibilities of cheating were checked. Twelve end-of-block exams were included, comprising four exams each for the first-, second-, and third-year students. Eight exams were conducted as PBEs (three for the first year, two for the second year, and three for the third year), and four were given as WBEs (one for each of first and third years, and two for the second year). Results of students who took all tests for a given level were included, whereas those of students who missed some exams were excluded. Table 1 shows the duration, year level, type of course exam, and number of students included in the study.

Statistical analysis

Data were analysed using SPSS version 21 (IBM Corp., Armonk, New York, USA) for Windows. The analysis included:

- A comparison of each batch's mean class score for PBEs and WBEs
- The correlation of each student's score in each exam, whether WBE or PBE
- A comparison of the scores between PBE and WBE among the top ten students and ten lowest-achieving students

Comparisons among multiple groups were made using the method of Šidák.[17] Data were analysed using difference-in-means tests, ANOVA, post-hoc Tukey's test, and the Pearson correlation test.

- $r = 0.00$ to 0.30 (0.00 to -0.30) was considered as negligible correlation
- $r = 0.30$ to 0.50 (-0.30 to -0.50) was considered as low positive (negative) correlation
- $r = 0.50$ to 0.70 (-0.50 to -0.70) was considered as moderate positive (negative) correlation
- $r = 0.70$ to 0.90 (-0.70 to -0.90) was considered as high positive (negative) correlation
- $r = 0.90$ to 1.00 (-0.90 to -1.00) was considered as very high positive (negative) correlation[18]

P values were considered for multiple comparisons. For all analyses, significance was established a priori as $P < 0.05$.

For the comparison of individual students' scores, we considered a difference of $\geq$5% as effective and that of $\geq$20% as high.

Results

For first-year students, the mean percentage scores were 69.11% ± 11.38%, 81.04% ± 9.28%, and 73.21% ± 12.86% on the PBEs and 75.29% ± 12.59% for the WBE (Table 2).

ANOVA revealed highly significant difference in the mean score among different tests (Table 2). Post-hoc Tukey's analysis showed that the score for PBE-2, the exam of the shortest course, was significantly higher than that of all other tests, whether PBE or WBE. The difference was significant among the three PBEs, and between the WBE and each of first and second PBEs, but it was not significant between the WBE-1 and PBE-3.

Table 3 shows that the correlation among the marks of individual students in the four exams was always positive. It is noted that the correlation among the three PBEs was high ($r = 0.782$–0.817), whereas it was moderate between the single WBE and each of the three PBEs ($r = 0.574$, 0.566, 0.626).

Table 1: Durations of courses, type of exams and number of students included in the study.

Course/Block	Year level	Duration in weeks	Exam type & order	Number of students included
Man and His Environment, and Metabolism	1st	8	PBE-1	131
Growth and Development		5	PBE-2	
Principles of Disease		7	PBE-3	
Musculoskeletal system		10	WBE-1	
Endocrine and Reproductive System	2nd	9	PBE-4	134
Hemopoitic and Immune Systems		8	PBE-5	
Cardiovascular system		8	WBE-2	
Respiratory System		8	WBE-3	
Digestive System	3rd	8	PBE-6	122
Urinary System		5	PBE-7	
Nervous System & special senses		10	PBE-8	
Integrated Multi-systems & therapeutics		10	WBE-4	

PBE: Paper-based exam.
WBE: Web -based exam.

Table 2: Percentage score of 1st year students (n = 131) in different exams.

	(ANOVA)					Post-hoc analysis	
	PBE-1	PBE-2	PBE-3	WBE-1	F value	Q value	Critical Range
Lowest mark %	34.0	52.0	35.0	35.0	23.95[a]	3.633	3.687
Highest mark %	93.0	96.0	96.0	97.5			
Mean mark % (SD)	69.11 (11.38)	81.04 (9.28)	73.21 (12.86)	75.29 (12.59)			

[a] P-value was significant at level of 0.001.

Table 3: Correlation coefficient (*r*) of marks of 1st year students in different exams.

	PBE-1	PBE-2	PBE-3	WPE-1
PBE-1	–			
PBE-2	0.817	–		
PBE-3	0.782	0.809	–	
WPE-1	0.574	0.566	0.626	–

A comparison of the average marks of each of the top 10 students on PBEs and their own marks in the WBE revealed that nine of them had a lower score on the WBE than on the PBEs. The decrease in score was ≥5% in five students, although the difference between their mean scores in PBEs and WBE was not statistically significant (Table 4). Furthermore, five out of the 10 lowest-achieving students had higher scores on the WBE than on the PBEs. Although

Table 4: Comparison of scores of the 1st-year high and low achievers on PBE and WBE of (n = 131).

Top 10 students			Bottom 10 students		
Student rank (based on average score in PBEs)	Average score on PBEs	Score on WBE-1	Student rank (based on average score in PBEs)	Average score on PBEs	Score on WBE-1
1	94.7 ± 1.5	92.5	131	47.0 ± 11.53	37.5[a]
2	90.7 ± 5.9	85.0[a]	130	47.7 ± 4.0	85.0[c]
3	90.7 ± 3.2	85.0[a]	129	48.7 ± 13.3	57.5[b]
4	90.3 ± 5.0	90.0	128	51.0 ± 6.6	35.0[a]
5	90.0 ± 2.7	85.0[a]	127	52.0 ± 4.4	45.0[a]
6	89.0 ± 0.0	97.5[b]	126	54.7 ± 11.6	47.5[a]
7	89.0 ± 4.9	85.0	125	55.7 ± 18.0	47.5[a]
8	88.7 ± 3.5	87.5	124	55.7 ± 10.6	60.0
9	88.3 ± 5.5	82.5[a]	123	56.7 ± 15.3	70.0[b]
10	88.3 ± 1.53	85.0	122	58.3 ± 6.1	80.0[c]
Mean score	90.0 ± 3.8	87.5	**Mean score**	52.7 ± 10.1	56.5
P = 0.140306			***P*** = 0.516567		

[a] Decrease of ≥5%.
[b] Increase of ≥5%.
[c] Increase of ≥20%.

Table 5: Percentage score of 2nd year students (n = 134) in different exams.

	PBE-4	PBE-5	WBE-2	WBE-3	F value (ANOVA)
Lowest mark %	37	31	25	25	1.282 (NS)
Highest mark %	94	93	97.5	90	
Mean mark % (SD)	69.35 (11.2)	70.17 (12.9)	69.31 (16.1)	67.11 (13.1)	

NS: no statistically significant difference.

Table 6: Correlation coefficient (*r*) of marks of 2nd year students in different exams.

	PBE-4	PBE-5	WBE-2	WBE-3
PBE-4	–			
PBE-5	0.803	–		
WBE-2	0.606	0.714	–	
WBE-3	0.547	0.641	0.696	–

two of those students had >20% increase in the score, the difference between the mean scores in the two test modalities was not statistically significant (Table 4).

Second-year students took two PBEs and two WBEs. The mean percentage marks were 69.35% ± 11.2% and 70.17% ± 12.9% for the PBEs and 69.31% ± 16.1% and 67.11% ± 13.1% for the WBEs. No statistically significant difference was found among the scores in the different tests (Table 5).

Table 9: Correlation coefficient (*r*) of marks of 3rd year students in different exams.

	PBE-6	PBE-7	PBE-8	WBE-4
PBE-6	–			
PBE-7	0.847	–		
PBE-8	0.842	0.827	–	
WBE-4	0.626	0.685	0.609	–

All exam scores of the second-year students showed positive correlation. The correlation was high between the two PBEs ($r = 0.803$) and two WBEs ($r = 0.696$). It was moderate between both of the PBEs and PBE-4 ($r = 0.606$, 0.547) and between PBE-5 and WBE-3 ($r = 0.641$).

Table 7: Comparison of scores of the 2nd-year high and low achievers on PBE and WBE of (n = 134).

Top 10 students			Bottom 10 students		
Student rank (based on average score in PBEs)	Average score on PBEs	Average score on WBEs	Student rank (based on average score in PBEs)	Average score on PBEs	Average score on WBEs
1	93.0 ± 1.4	93.8 ± 5.3	134	34.5 ± 4.9	33.8 ± 12.4
2	92.5 ± 0.7	82.5 ± 0.0[a]	133	36.0 ± 1.4	55.0 ± 0.0[b]
3	89.5 ± 0.7	86.3 ± 5.3	132	43.5 ± 0.7	40.0 ± 14.1
4	89.0 ± 1.4	90.0 ± 3.5	131	44.5 ± 0.7	43.8 ± 15.9
5	88.0 ± 1.4	90.0 ± 0.0	130	45.0 ± 4.2	65.0 ± 3.5[c]
6	85.5 ± 4.9	75.0 ± 3.5[a]	129	45.0 ± 8.5	56.3 ± 1.8[b]
7	84.5 ± 6.4	88.8 ± 5.3	128	46.5 ± 6.4	50.0 ± 10.6
8	84.5 ± 3.5	88.8 ± 1.8	127	47.0 ± 0.0	70.0 ± 3.5[c]
9	84.5 ± 3.5	85.0 ± 14.1	126	47.0 ± 1.4	46.3 ± 1.8
10	84.5 ± 3.5	82.5 ± 7.1	125	47.0 ± 11.3	50.0 ± 3.5
Mean score	87.6 ± 2.8	86.3 ± 4.6	**Mean score**	43.6 ± 4.0	51.0 ± 6.7
P = 0.522428			***P*** = 0.073795		

[a] Decrease of ≥5%.
[b] Increase of ≥5%.
[c] Increase of ≥20%.

Table 8: Percentage score of 3rd-year students (n = 122) in different exams.

	(ANOVA)					Post-hoc analysis	
	PBE-6	PBE-7	PBE-8	WBE-4	F value	Q value	Critical Range
Lowest mark %	36	51	29	37.5	22.141[a]	3.633	4.195
Highest mark %	94	96	95	92.5			
Mean mark % (SD)	71.84 (12.1)	78.03 (11.6)	67.96 (14.2)	65.59 (13.0)			

[a] P-value was significant at level of 0.001.

Table 10: Comparison of scores of the 3rd-year high and low achievers on PBE and WBE of (n = 122).

Top 10 students			Bottom 10 students		
Student rank (based on average score in PBEs)	Average score on PBEs	Score on WBE-4	Student rank (based on average score in PBEs)	Average score on PBEs	Score on WBE-4
1	94.7 ± 0.6	90.0	122	39.3 ± 12.4	57.5[c]
2	91.7 ± 2.5	85.0[a]	121	43.3 ± 7.5	45
3	91.7 ± 2.1	77.5[a]	120	47.3 ± 12.9	45
4	91.3 ± 2.5	90.0	119	51.3 ± 7.8	45[a]
5	91.3 ± 2.1	90.0	118	51.7 ± 10.8	77.5[d]
6	91.3 ± 2.9	92.5	117	51.7 ± 4.9	45[a]
7	89.7 ± 1.5	62.5[b]	116	53.0 ± 2.7	52.5
8	89.3 ± 3.1	77.5[a]	115	53.7 ± 5.7	52.5
9	88.3 ± 3.2	77.5[a]	114	54.0 ± 9.2	62.5[c]
10	88.0 ± 1.7	70.0[a]	113	54.3 ± 7.2	50
Mean score	90.7 ± 2.2	81.3	**Mean score**	50.0 ± 8.1	53.3
P = 0.014805[e]			***P*** = 0.385943		

[a] Decrease of ≥5%.
[b] Decrease of ≥20%.
[c] Increase of ≥5%.
[d] Increase of ≥20%.
[e] Statistically significant the current.

However, it was high between PBE-5 and WBE-2 ($r = 0.714$; Table 6).

Among the top 10 s-year students, two had an average score in the WBEs that was >5% lower than that in the PBEs. Likewise, two of the 10 lowest-achieving students had scores on the WBE that were ≥5% than those on the PBEs, whereas another two had ≥20% higher scores. For both high and low achievers, no statistically significant difference was found in the mean scores between the PBEs and WBEs (Table 7).

Table 8 shows the descriptive statistics of the performance of third-year students in the three PBEs and one WBE conducted during the year. The variation of scores in the different exams was evident. The mean score for PBE-7—the course with a short duration—was statistically significantly higher than that for the three other exams. However, the mean score for WBE-4 was statistically significantly lower than that of PBE-6 and PBE-7.

The performance of an individual student on different exams showed a positive correlation. The correlation was high between the three PBEs ($r = 0.847$, 0.842, 0.827) and between PBE-7 and the conducted WBE ($r = 0.685$) but was moderate between the WBE and PBE-6 ($r = 0.626$) and PBE-8 ($r = 0.609$; Table 9).

In nine out of the of the top 10 students, the marks in the WBE were lower than the average marks on the PBEs (i.e., a ≥5% decrease in five students and a ≥20% decrease in four students). The difference between their mean scores in the PBEs and WBE was statistically significant ($P \leq 0.05$; Table 10). For the low achievers, two had a decrease of ≥5%, two increased by ≥5%, and one had >20% increase in the scores. The difference between the mean scores in the two test modalities was not statistically significant (Table 10).

Discussion

The COVID-19 pandemic affected almost every aspect of life. Social distancing and lockdown to minimise the transmission of the virus forced the immediate worldwide transition to distance learning and remote assessments. The current study was designed to evaluate the effectiveness of WBEs, to test whether their outcomes were different from those of PBEs, and to assess the reliability of WBEs.

The current study revealed variations in the scores of students in three batches in different exams conducted during the academic year 2019/2020, whether PBEs or WBEs. Statistically significant differences were found in some instances.

A variation in the score of the same batch in different tests is frequently reported. Reports included various factors, such as course contents, course duration, the nature of the assessment tool, and students' motivation, as causes for those differences.[19–22] In the current study, the PBE scores in courses with short durations (five weeks) were significantly higher than those in courses with long durations. The shorter content of courses with shorter duration is the probable cause.

In this study, no specific trend/pattern was found for the difference between PBEs or WBEs. The mean score of WBE for the third-year students was lower than the scores for PBEs, whereas for the first- and second-year students, the results of WBEs was either lower or higher than PBEs. This matches the results of previous studies that indicated no difference between the results of computer-based assessments (CBAs) and paper-based assessments (PBAs) in personality assessments,[23] in the reading abilities of students up to the 12th grade,[24] and in exams in higher education.[25]

The significantly lower score in the last course in the preclinical phase (WBE-4) compared to that in other courses for the same batch was consistent with the results of Reed and Holley,[26] who showed that throughout sequential accounting courses, students' grades tend to drop as they progress in the programme. In the spiral approach adopted in QUCOM, later courses contain more complex

and integrated information than previous ones in the phase, and tests target increasing levels of cognition. The attribution of low scores on tests of complexity and difficulty has been previously reported.[27] In the current study, PBEs and WBEs were type-A MCQs in nature. Additionally, students became familiar with online tests through mock tests and quizzes. Thus, the possibility of online assessment to be the reason for the observed decline was greatly minimised. Given that the score of the last course did not differ much from that of the preceding one (PBE-8) and that both courses were 10-weeks in duration, long course durations tend to be associated with low scores. The lower score may be attributed to the stress caused by the COVID-19 pandemic; reports showed that stress caused by out-of-class circumstances influences student performance.[28,29]

However, the number of test items varied. Each PBE contained 100 items with 120 min in duration (72 s per item), whereas each WBE comprised 40 items with 50 min in duration (75 s per item). WBEs may be expected to yield a better outcome than PBEs because of the lower number of test questions. Although an association between a long test and difficulty was reported by Alamro (2019),[30] it was not observed in the current study.

Further analysis revealed that the correlation of an individual student's marks on different exams was consistently positive. Among PBEs, the correlation was always high ($r = 0.782, 0.847$) and was slightly high for a single situation of two WBEs for the second-year students ($r = 0.696$). However, the correlation between scores on PBEs and WBEs was moderate in most cases and slightly high in two cases ($r = 0.685, 0.714$).

The observed variance in correlation may be attributed to differences in the attitude of students towards the different test modalities. In the current study, top students tended to have lower marks on WBEs than on PBEs, whereas low performers generally obtained higher scores on WBEs than on PBEs. Karay et al. reported that low performers guess significantly more in computer-based exams than in paper–pencil tests, although no explanation could be inferred.[31]

The decline of scores of the third-year top achievers in WBE-4 was consistent with the decrease of the mean class score in the test. However, the possibility of malpractice as the cause of the improved performance of the low performers cannot be ignored. Among 30 low-achieving students in the three batches, 11 (four in each of the first and second years and three in the third year) had increased WBE scores by $\geq$5%, and almost half of them (two in each of first and second years and one in the third year) had an increased score by $\geq$20%. The difference was not statistically significant for any of the three batches as it may have been concealed by intra-group variance, given that five first-year students and two third-year students showed decreased scores by $\geq$5%.

Malpractices during online examinations is a key concern for many health profession educators.[13] Approaches to minimise plagiarism included reliance on students' commitment to professionalism and timed/locked down open-book examinations; furthermore, replacing exams with remotely-completed project work is an alternative.[13] A model depending on the arrangement of items to prevent cheating in CBAs was also suggested.[32]

In the current study, measures to prevent cheating included dilution of weight of final exam through distribution of the assessment tasks throughout the entire period of distance learning, with less weight for the end-of-course exam. Exams comprising type-A MCQs to ensure that answers are vignette-dependent and the lead–in alone provides no clue is another factor. Additionally, questions were randomly ordered for different students, and time-locking and disabling the go back function were also used to prevent cheating through group answering. The possibility of a person taking the exam on the behalf of another is still present because the use of webcam to invigilate was prohibited owing to cultural constraints and university regulations preventing viewing of girls.

The weight dilution of individual assessment tasks could help motivate the learning of students and increase their commitment to professionalism.[33] In the present study, the $\geq$20% increase in scores of five out of 30 low achievers raises queries about the validity of this approach in preventing plagiarism. Furthermore, relying on trust in the probity pledges of students is not always effective.[13]

Using vignette-dependent MCQs can eliminate the possibility of finding answers online through posting the lead–in. Time-locking minimises the possibility of analysing a scenario by searching for its key issues online. Although this manoeuvre would negate the possibility of finding answers on the internet, but it cannot prevent another person from taking a test as proxy for a student.

The present study revealed no remarkable difference in the outcomes of PBEs and WBEs targeting the cognitive domain, which is consistent with previous reports on the similarities of assessment results between CBAs and PBAs.[23–25] This finding justifies the use of remote assessments for the evaluation of a student's achievement of learning outcomes. However, the results of the current study cannot support the use of online assessments for decision-making tests as the probability of plagiarism cannot be ignored. Many medical educationists are concerned about academic malpractices during online multiple mini-interviews for admission and high-stakes tests.[13]

WBEs would be beneficial for formative, and continuous low-stakes assessments. In addition to low copying and printing costs, immediate feedback can be provided to students, especially in MCQ tests—this will enhance their learning.[34] This method can aid large-scale, multi-institutional, real-time assessments. The nationwide use of this approach can ensure the unification of assessments for a given content and timely feedback to students and educators. In the case of other medical colleges with similar intended learning outcomes (ILOs), such as those of SaudiMEDs in KSA,[35,36] unified, trans-institutional tests can provide efficient assessment processes. A high-quality online test can offer uncompromised educational outcomes.[32] The unification of assessment through WBEs can increase an item writing workforce and allow rapid, regular change of questions and topics; furthermore, students can be assessed on their real learning, rather than rote learning.[33] The adoption of this approach is ideal for formative progress testing.

Compulsory social distancing because of the COVID-19 pandemic made online learning and testing practices popular. Today, teaching can be conveniently and professionally

performed during lockdowns. Continuing with distance learning after the pandemic will reduce the need for conventional infrastructure, facilitate worldwide sharing of teaching resources, and help deliver high-quality education.[32]

Online assessment, unlike online teaching, does not enjoy global acceptance. Although its validity for high-stakes exams is questionable, its benefits justify its use in low-stakes exams and formative assessment.[25] Young people—the digital natives—would surely be interested in CBAs; therefore investing in CBAs could be beneficial for them.[25] Furthermore, online assessments must be revisited, and the inclusion of assessment modalities and measures to avoid academic malpractice should be considered.

The authors are aware that the present study was conducted in one institution with a limited number of students. The inclusion of experiences of other universities with remote assessment will enrich knowledge about the present study's pros and cons, and deepen our understanding of the subject area. Further studies are recommended before any action based on these results is taken.

Conclusion

The present study showed the similarity between students' results for WBEs and PBEs, both on class average and for individual students' scores. Although there were some variations among results of the two approaches of assessment, there was no special trend in those variations as the results for the WBEs were sometimes higher and sometimes lower than those for PBEs. This justifies the consideration of online assessment along with the increasing tendency of distance learning. However, the possibility of academic malpractices during online exams could not be completely negated in the present study, and a few low-achievers may exploit the non-strict monitoring to their advantage.

Recommendation

WBEs are an ideal approach for formative assessment, progress testing, and frequent low-weight continuous assessment. In case WBEs are used for high-stakes tests, additional measures against academic misconduct are required to maximise the validity of assessment.

Source of funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors have no conflicts of interest to declare.

Ethical approval

The authors confirm that this study has been conducted in accordance with COPE rules and regulations. Given the nature of the study, the IRB review was not required.

Authors' contributions

MNEDS initiated the research idea, interpreted the data, and wrote and edited the manuscript. TARS entered, analysed, and tabulated the data. MMW and ASA finalised the manuscript, inserted citations, wrote the abstract and introduction, and prepared the manuscript for publication. All authors have critically reviewed and approved the final draft and are responsible for the content and similarity index of the manuscript.

References

1. UNESCO. *COVID-19 Educational Disruption and Response*; 2020. Available at: https://en.unesco.org/themes/education-emergencies/coronavirus-school-closures. [Accessed 8 December 2020].
2. WHO. *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*; 2020. Available at: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020. [Accessed 8 December 2020].
3. Ashokka B, Ong SY, Tay KH, Loh NHW, Gee CF, Samarasekera DD. Coordinated responses of academic medical centres to pandemics: sustaining medical education during COVID-19. **Med Teach 2020**: 1–10.
4. UN. *Education during COVID-19 and beyond*; 2020. Available at: https://www.un.org/development/desa/dspd/wp-content/uploads/sites/22/2020/08/sg_policy_brief_covid-19_and_education_august_2020.pdf. [Accessed 8 December 2020].
5. Hafiz H, Oei S-Y, Ring DM, Shnitser N. *Regulating in pandemic: evaluating economic and financial policy responses to the coronavirus crisis*. Boston College Law School Legal Studies Research Paper; 2020 (527).
6. Taha M, Abdalla M, Wadi M, Khalafalla H. Curriculum delivery in Medical Education during an emergency: a guide based on the responses to the COVID-19 pandemic. **MedEdPublish 2020**; 9(1): 69.
7. Ahmed H, Allaf M, Elghazaly H. *COVID-19 and medical education*. The Lancet Infectious Diseases; 2020.
8. Rose S. Medical student education in the time of COVID-19. **J Am Med Assoc 2020**; 323(21): 2131–2132.
9. Wadi M, Abdalla ME, Khalafalla H, Taha MH. The assessment clock: a model to prioritize the principles of the utility of assessment formula in emergency situations, such as the COVID-19 pandemic. **MedEdPublish 2020**; 9.
10. Khan RA, Jawaid M. Technology enhanced assessment (TEA) in COVID 19 Pandemic. **Pak J Med Sci 2020**; 36(COVID19-S4): S108.
11. Dennick R, Wilkinson S, Purcell N. Online eAssessment: AMEE guide No. 39. **Med Teach 2009**; 31(3): 192–206.
12. Khalil R, Mansour AE, Fadda WA, Almisnid K, Aldamegh M, Al-Nafeesah A, et al. The sudden transition to synchronized online learning during the COVID-19 pandemic in Saudi Arabia: a qualitative study exploring medical students' perspectives. **BMC Med Educ 2020**; 20(1): 285.
13. Cleland J, McKimm J, Fuller R, Taylor D, Janczukowicz J, Gibbs T. Adapting to the impact of COVID-19: sharing stories, sharing practice. **Med Teach 2020**: 1–4.
14. Guidance MOE. *Manual for arrangement of testing and evaluation during COVID-19 lockdown*. Ministry of Education; 2020.
15. Huang IA, Lu Y, Wagner JP, Quach C, Donahue TR, Tillou A, et al. Multi-institutional virtual mock oral examinations for general surgery residents in the era of COVID-19. **Am J Surg 2021**; 221(2): 429–430.

16. Elzainy A, El Sadik A, Al Abdulmonem W. *Experience of e-learning and online assessment during the COVID-19 pandemic at the College of Medicine*. Qassim University. Journal of Taibah University Medical Sciences; 2020.
17. Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. **J Am Stat Assoc 1967**; 62(318): 626–633.
18. Jaadi Z. Everything you need to know about interpreting correlations. (available at: https://towardsdatascience.com/eveything-you-need-to-know-about-interpreting-correlations-2c485841c0b8), accessed on 26 July 2021.
19. Wass V, McGibbon D, Van der Vleuten C. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? **Med Educ 2001**; 35(4): 326–330.
20. Willingham WW, Pollack JM, Lewis C. Grades and test scores: accounting for observed differences. **J Educ Meas 2002**; 39(1): 1–37.
21. Pomplun M, Ritchie T, Custer M. Factors in paper-and-pencil and computer reading score differences at the primary grades. **Educ Assess 2006**; 11(2): 127–143.
22. Nair BR, Moonen-van Loon JM, Parvathy M, Jolly BC, van der Vleuten CP. Composite reliability of workplace-based assessment of international medical graduates. **Med J Aust 2017**; 207(10): 453.
23. Ford BD, Vitelli R, Stuckless N. The effects of computer versus paper-and-pencil administration on measures of anger and revenge with an inmate population. **Comput Hum Behav 1996**; 12(1): 159–166.
24. Wang S, Jiao H, Young MJ, Brooks T, Olson J. Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: a meta-analysis of testing mode effects. **Educ Psychol Meas 2008**; 68(1): 5–24.
25. Boevé AJ, Meijer RR, Albers CJ, Beetsma Y, Bosker RJ. Introducing computer-based testing in high-stakes exams in higher education: results of a field experiment. **PLos One 2015**; 10(12):e0143616.
26. Reed SA, Holley JM. The effect of final examination scheduling on student performance. **Issues Account Educ 1989**; 4(2): 327–344.
27. Anakwe B. Comparison of student performance in paper-based versus computer-based testing. **J Educ Bus 2008**; 84(1): 13–17.
28. Sakka S, Nikopoulou VA, Bonti E, Tatsiopoulou P, Karamouzi P, Giazkoulidou A, et al. Assessing test anxiety and resilience among Greek adolescents during COVID-19 pandemic. **J Mind Med Sci 2020**; 7(2): 173–178.
29. Husky MM, Kovess-Masfety V, Swendsen JD. Stress and anxiety among university students in France during Covid-19 mandatory confinement. **Compr Psychiatr 2020**; 102: 152191.
30. Alamro AS. The effect of order of mcq items on difficulty index. **Int J Med Sci Educ 2019**; 6(2): 1–9.
31. Karay Y, Schauber SK, Stosch C, Schüttpelz-Brauns K. Computer versus paper—does it make any difference in test performance? **Teach Learn Med 2015**; 27(1): 57–62.
32. Nizam NI, Gao S, Li M, Mohamed H, Wang G. Scheme for cheating prevention in online exams during social distancing. **Preprints 2020**. https://doi.org/10.20944/preprints202004.0327.v1.
33. Bloxham S, Boyd P. *Developing effective assessment in higher education: a practical guide: a practical guide*. UK): McGraw-Hill Education; 2007.
34. Karay Y, Schauber SK, Stosch C, Schuettpelz-Brauns K. Can computer-based assessment enhance the acceptance of formative multiple choice exams? A utility analysis. **Med Teach 2012**; 34(4): 292–296.
35. SaudiMEDs. *SaudiMEDs framework (Saudi Medical Education Directives Framework)*. Education and Training Evaluation Commission.; 2017. Available at: https://etec.gov.sa/en/productsandservices/NCAAA/AccreditationProgrammatic/Pages/Medical-Colleges.aspx. [Accessed December 2020].
36. Tekian AS, Al Ahwal MS. Aligning the SaudiMED framework with the national Commission for academic Accreditation and assessment domains. **Saudi Med J 2015**; 36(12): 1496.

How to cite this article: Saleh MN, Salem TAR, Alamro AS, Wadi MM. Web-based and paper-based examinations: Lessons learnt during the COVID-19 pandemic lockdown. J Taibah Univ Med Sc 2022;17(1):128–136.