Creative Education, 2019, 10, 1813-1818
http://www.scirp.org/journal/ce
ISSN Online: 2151-4771
ISSN Print: 2151-4755

The Effect of Scrambling Test Item on Students' Performance and Difficulty Level of MCQs Test in a College of Medicine, KKU

Ismail Satti[1], Bahaeldin Hassan[2*], Abdulaziz Alamri[3], Muhammad Abid Khan[3], Ayyub Patel[4]

[1]Department of Obstetrics & Gynecology, College of Medicine, King Khalid University, Abha, KSA
[2]Department of Medical Education and Obstetrics & Gynecology, College of Medicine, King Khalid University, Abha, KSA
[3]Department of Medical Education, College of Medicine, King Khalid University, Abha, KSA
[4]Department of Biochemistry and Medical Education, King Khalid University, Abha, KSA
Email: *bahasuikt@hotmail.com

How to cite this paper: Satti, I., Hassan, B., Alamri, A., Khan, M. A., & Patel, A. (2019). The Effect of Scrambling Test Item on Students' Performance and Difficulty Level of MCQs Test in a College of Medicine, KKU. *Creative Education, 10,* 1813-1818.
https://doi.org/10.4236/ce.2019.108130

Received: July 8, 2019
Accepted: August 9, 2019
Published: August 12, 2019

Abstract

Background: Multiple Choice Tests (MCQs) are commonly used assessment tool in medical schools, which is delivered to our student in four versions (A, B, C and D) to avoid cheating. The aim of this study was to investigate the effect of scrambling test items on students' performance and the difficulty level of each version, so as to decide on continuing randomization of test items or keeping it without randomization. **Methods:** A prospective, cross-sectional study was conducted, the participants were the 5th year undergraduate medical students during their major course of obstetrics and gynecology. Three tests where items were randomized are delivered to the students. After correction, the marks obtained by the candidates and difficulty index of each version were entered into the Statistical Package for Social Sciences (SPSS) version 20 and comparison amongst these four versions was carried out through analysis of variance (ANOVA). A p-value of <0.05 was considered as statistically significant. **Results:** No significant difference was found in the mean difficulty index for different versions in each test and there are no statistically different results when we compared version A mean students' scores to other versions (B, C and D) after applying ANOVA analysis (F = 1.14, p = 0.42), (F = 0.75, p = 0.69) and (F = 1.29, p = 0.34); (F = 0.84, p = 0.62), (F = 0.81, p = 0.64) and (F = 0.62, p = 0.79); (F = 0.62, p = 0.79), (F = 0.35, p = 0.95) and (F = 0.83, p = 0.64) for test 1, 2 and 3 respectively. **Conclusions:** Our study revealed that randomization of MCQs test item into versions to avoid cheating does not affect student performance or the difficulty level of the exam.

Keywords

MCQs Test, Scrambling, Version, Student's Score, Difficulty Index

1. Introduction

The primary goal of any assessment of students is to provide valid and reliable evaluations of students' knowledge and skills as well as provision of accurate feedback to students about their performance (Al Mahmoud, Elzubeir, Shaban, & Branicki, 2015). We are increasingly dependent on Multiple Choice Tests (MCQs) as the sole tool for assessment because it is valid, objective and cost effective tool of assessment.

Examination malpractices are acts that contravene the rules and regulations which govern the conduct of examinations (Ollennu, 2015).

Randomization of test item into different versions that includes the same questions can minimize the chance of cheating by students, while keeping the level of difficulty of the exam constant across students since every version contains the same questions (Sue, 2009).

Several studies have suggested that changing the position of an item on an operational exam relative to its position during trial testing development leads to a change in the difficulty of the item (Schroeder, Murphy, & Holme, 2012).

The difficulty index, symbolized as p, can range from 0 (no one selected the keyed option) to 1.00 (everyone selected it). Naturally, overall test scores tend to be higher when the items on a test have higher p values, and vice versa (Di Battista & Kurzawa, 2011).

In our institution (King Khalid University, Faculty of Medicine, Saudi Arabia), randomization of test item into four versions (A, B, C and D) of the same test is done to avoid cheating and this is a mandatory requirement of every test before approval by the academic office.

Each version contains the same test items in a different order. In version (A), questions were ordered according to the coverage of the course materials in the class, in version (D) the questions were ordered in reverse sequences to version (A) and versions (B) and (C) were randomized. We observe that students who took version (A) finish the exam and collect their papers earlier than other versions.

This study was conducted to investigate the effect of scrambling the test questions on student performance and difficulty index of each test version. The difficulty index of an item is the proportion of examinees who selected the keyed option.

2. Methods

A prospective, cross-sectional study was carried out in College of Medicine-King Khalid University-Saudi Arabia, participants were 5th year undergraduate medical students who completed their major course of obstetrics and gynecology, the whole course duration is 8 weeks, after course blueprint, three tests of single best answer types were designed by teachers who taught the course at week 4, 6 and 8 during the second semester of academic year 2017-2018.

There were four versions of each exam. In the version (A), multiple choice questions were ordered according to material coverage in the class. In versions

(B and C), multiple choice questions were placed in random order, that is, unrelated to the order that the material was taught in the class. In version (D), multiple choice questions were placed in reverse order to version (A).

Ninety eight (98) undergraduate medical students were divided randomly into four versions of each test (A, B, C and D) every time they sat for the exam. Post-test item analysis was conducted for each exam and average student's score for each version was calculated, in addition to difficulty index of each version of the three exams. The marks obtained by the candidates and difficulty index of each version were entered into the Statistical Package for Social Sciences (SPSS) version 20 and comparison amongst the marks of candidates in these four versions were carried out through analysis of variance (ANOVA). A p-value of < 0.05 was considered as statistically Significant.

3. Results & Discussion

Difficulty level of the versions in each test was recorded from post-test item analysis, it is reflected as mean difficulty index in Table 1, for test 1 (0.69, 0.72, 0.69 and 0.68), test 2 (0.65, 0.66, 0.70 and 0.68) and test 3 (0.78, 0.78, 0.75 and 0.73) for versions A, B, C and D respectively. Version comparison was done in each test through Analysis of Variance (ANOVA). No significant difference was found in the mean difficulty index for different versions in each test. Results presented in Table 2 are shown ($F = 0.99$, $p = 0.49$), ($F = 1.50$, $p = 0.16$) and ($F = 1.46$, $p = 0.17$) for comparison of version A to B, A to C and A to D respectively in test 1 and similar non-significant results were obtained when comparing the versions B, C and D to version A in test 2 and 3.

Table 3 showed the average students' scores in each versions of the three tests (1, 2 out of 50 marks and 3 out of 60 marks), (34.9, 36, 34.5 and 34), (32.1, 33.1, 34.3 and 43.9) and (46.8, 47, 45.2 and 43.9) for versions A, B, C and D respectively.

Again there are no statistically different results when we compared version A mean students' scores to other versions (B, C and D) after applying ANOVA analysis in all three tests as showed in Table 4. ($F = 1.14$, $p = 0.42$), ($F = 0.75$, $p = 0.69$) and ($F = 1.29$, $p = 0.34$); ($F = 0.84$, $p = 0.62$), ($F = 0.81$, $p = 0.64$) and ($F = 0.62$, $p = 0.79$); ($F = 0.62$, $p = 0.79$), ($F = 0.35$, $p = 0.95$) and ($F = 0.83$, $p = 0.64$) for test 1, 2 and 3 respectively.

Table 1. Descriptive analysis of difficulty index among three MCQs tests.

Test 1				Test 2				Test 3			
N	Exam Versions	Mean Difficulty Index	S.D	N	Exam Versions	Mean Difficulty Index	S.D	N	Exam Versions	Mean Difficulty Index	S.D
50	Version A	0.69	0.27	50	Version A	0.65	0.24	50	Version A	0.77	0.2
50	Version B	0.721	0.26	50	Version B	0.66	0.25	50	Version B	0.78	0.21
50	Version C	0.69	0.25	50	Version C	0.7	0.23	50	Version C	0.75	0.19
50	Version D	0.68	0.26	50	Version D	0.68	0.2	50	Version D	0.73	0.23

N = Number of Test Items. SD = Standard Deviation.

Table 2. ANOVA comparisons of difficulty index among the three MCQs test.

Test 1			Test 2			Test 3		
	F	Sig.		F	Sig.		F	Sig.
Version A & Version B	0.986	0.493	Version A & Version B	1.519	0.149	Version A & Version B	0.472	0.943
Version A & Version c	1.495	0.161	Version A & Version c	1.46	0.171	Version A & Version c	0.749	0.722
Version A & Version D	1.463	0.173	Version A & Version D	0.886	0.601	Version A & Version D	0.534	0.9

Level of significance is 5%.

Table 3. Descriptive analysis for the students' scores among the three MCQs tests.

Test 1				Test 2				Test 3			
N	Exam Versions	Mean Student Scores	S.D	N	Exam Versions	Mean Student Scores	S.D	N	Exam Versions	Mean Student Scores	S.D
24	Version A	34.9	4.2	24	Version A	32.1	6.2	24	Version A	46.8	7.7
24	Version B	36	4.5	24	Version B	33.1	6.1	24	Version B	47	4.98
24	Version C	34.5	5.1	24	Version C	34.3	5.5	24	Version C	45.2	9.4
26	Version D	34	4.08	26	Version D	46.8	6.7	26	Version D	43.9	8.9

N = Number of Students per version. SD = Standard Deviation.

Table 4. ANOVA comparisons of students' scores among the three MCQs test.

Test 1			Test 2			Test 3		
Version A Version B	1.143	0.416	Version A Version B	0.837	0.619	Version A Version B	0.617	0.79
Version A Version c	0.745	0.69	Version A Version c	0.81	0.64	Version A Version c	0.35	0.95
Version A Version D	1.29	0.34	Version A Version D	0.618	0.79	Version A Version D	0.831	0.64

4. Discussion

This study is the first study comparing more than one version of scrambled but similar-content MCQ papers in a medical school in Saudi Arabia.

Our study failed to identify any differences in the scores of students taking the version which followed content coverage sequence (version A), from other randomized versions (B, C and D).

Similar to our study, Sue D.L. concluded that, the technique of scrambling multiple-choice questions in order to reduce the benefits of student cheating during the exam can be done without risk of biasing student performance (Sue, 2009).

Another study in medical school in Pakistan comparing more than one sequence of scrambled but similar-content MCQ papers in a high-stake entrance

examination over 3 years from 2008 to 2011. It failed to identify any differences in the scores of students receiving the papers which followed content coverage sequence, from those that did not (Khan, Tabasum, Mukhtar, & Iqbal, 2013).

Zaman et al. concluded that item difficulty is not affected by the sequence of items in the test (Zaman, Niwaz, Faize, & Dahar, 2010).

On other hand, some studies have shown that there was indeed statistically significant difference in performance when the positions of the items were altered (Ollennu, 2015; Doerner & Calhoun, 2009; Raux, Sangnier, & Ypersele, 2017).

Although English is foreign language to our students, the sequence of items did not affect their performance, these findings contrast the results of Soureshjani K.H., who revealed that the sequence of items affect foreign language learners' performance (Soureshjani, 2011).

5. Conclusion & Recommendations

Up to our knowledge this is the first study comparing more than one version of scrambled but similar-content MCQ papers in a medical school in Saudi Arabia. Our study revealed that randomization of test item into versions to avoid cheating does not affect student performance or the difficulty level of the exam. Our institution can carry on their regulations of scrambling questions into different versions without hesitation. Further studies are recommended in this field to ensure better assessment of our students.

Conflicts of Interest

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

References

Al Mahmoud, T., Elzubeir, M., Shaban, S., & Branicki, F. (2015). An Enhancement-Focused Framework for Developing High Quality Single Best Answer Multiple Choice Questions. *Education for Health, 28,* 194-200. https://doi.org/10.4103/1357-6283.178604

Di Battista, D., & Kurzawa, L. (2011). Examination of the Quality of Multiple-Choice Items on Classroom Tests. *Canadian Journal for the Scholarship of Teaching and Learning, 2,* 4. https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Doerner, W., & Calhoun, J. (2009). *The Impact of the Order of Test Questions in Introductory Economics*. https://doi.org/10.2139/ssrn.1321906

Khan, J. S., Tabasum, S., Mukhtar, O., & Iqbal, M. (2013). The Effect on Student Performance of Scrambling Questions and Their Stems in Medical Colleges Admission Tests. *Journal of the College of Physicians and Surgeons Pakistan, 23,* 904-906.

Ollennu, S. N. N. (2015). The Impact of Item Position in Multiple-Choice Test on Student Performance at the Basic Education Certificate Examination (BECE) Level: University of Cape Coast. *Universal Journal of Educational Research, 3,* 718-723. https://doi.org/10.13189/ujer.2015.031009

Raux, M., Sangnier, M., & Van Ypersele, T. (2017). Scrambled Questions Penalty in Multiple Choice Tests: New Evidence from French Undergraduate Students. *Economics*

Bulletin, 37, 347-351.

Schroeder, J., Murphy, K. L., & Holme, T.A. (2012). Investigating Factors That Influence Item Performance on ACS Exams. *Journal of Chemical Education, 89,* 346-350. https://doi.org/10.1021/ed101175f

Soureshjani, K. H. (2011). Item Sequence on Test Performance: Easy Items First? *Language Testing in Asia, 1,* 46-59. https://doi.org/10.1186/2229-0443-1-3-46

Sue, D. L. (2009). The Effect of Scrambling Test Questions on Student Performance in a Small Class Setting. *Journal for Economic Educators, 9,* 32-41.

Zaman, A., Niwaz, A., Faize, F., & Dahar, M. (2010). Analysis of Multiple Choice Items and the Effect of Items' Sequencing on Difficulty Level in the Test of Mathematics. *European Journal of Social Sciences, 17,* 61-67.