International Journal of Medical Science and Education

www.ijmse.com

Original Research Article

pISSN- 2348 4438 | eISSN-2349- 3208

THE EFFECT OF ORDER OF MCQ ITEMS ON DIFFICULTY INDEX

Ahmad Saleh Alamro

1. Medical Education Department, College of Medicine, Qassim University

*Corresponding author – **Ahmad Saleh Alamro**
Email id – ah.alamro@qu.edu.sa

Received: 02/06/2019 ***Revised:10/06/2019*** ***Accepted: 18/06/2019***

ABSTRACT

Background: Multiple choice questions (MCQs) are substantially used to assess cognition in medical education worldwide. One of the limitations of MCQ tests is the ease of cheating. Use of a multi-version test can overcome this challenge because it makes cheating difficult while maintaining the same content to ensure fairness. The aims of the current study are to explore the effect of scrambling MCQ items on the difficulty index (DI) and to determine the subsequent effect on students' performance **Materials and Methods:** Six MCQ tests with different versions were taken by 514 third-year medical students who were in the preclinical phase at the College of Medicine, Qassim University. Each test comprised 100 A-type MCQs. Each exam was divided into quintuples, and the mean DI for each quintuple was calculated and analysed. **Results:** Generally, there was no difference in the mean DI for the different versions of the same test. The DI of the fourth and fifth quintuples was significantly lower than the first three ($p < 0.01$). This change was less evident for students with a more advanced level in the programme. **Conclusion:** The scrambling of MCQs on a multi-version test does not affect the test's overall difficulty. However, items may yield a lower DI when they are presented late in the test. The inclusion of more test questions may cause some difficulties regarding the length of the test rather than item quality being the cause.

Keywords: MCQ, item order, Difficulty Index, student performance

INTRODUCTION

Medical schools require a high-quality assessment system because tests provide important feedback to stakeholders and educators regarding students' educational progress. There is a growing body of literature that addresses the importance of validity, reliability and other quality indicators of assessments **(1–4)**. Different methods are used to assess students' competencies in medical education. Among them, multiple choice question (MCQ) tests are used substantially worldwide **(5)**. MCQs are used for both low-stakes and high-stakes exams because they provide broad sampling, they can assess a large number of examinees, they produce timely results and they are feasible **(6, 7)**.

A major concern with MCQ test construction is ensuring reliability **(8)**, which depends on grading consistency and discrimination among students of differing performance levels. Reliability is influenced by many factors, such the test's length, time limit, appropriate sampling and difficulty **(9–11)**.

Post-exam psychometric analyses are used as quality indicators of a test and its particular items **(12, 13)**. Several studies have reported that high-quality, credible MCQs will result in better validity and higher reliability of an assessment **(6, 8, 14–16)**.

One limitation of MCQ tests is the ease of cheating **(17, 18)**. The use of a multi-version test can overcome this challenge because it makes cheating difficult while maintaining the same content to ensure fairness**(19)**.

The Qassim University College of Medicine uses multi-version MCQ tests with other modalities to assess students' achievement at the completion of different courses.

The purposes of this study were to explore the effect of scrambling MCQ items on the difficulty index (DI) and to determine the subsequent effect on students' performance.

MATERIALS AND METHODS

The Qassim College of Medicine uses a 100 A-Type MCQ test at the end of a block (EOB) in the preclinical phases (the first, second and third year) to assess students' learning status. These exams are multi-versions that use computerised random scrambling. The data in this study were extracted from an item analysis of 6 EOB exams, which included 600 questions. The exams included two for first-year students, two for second-year students and two for third-year students. This study was done as part of an internal quality assurance assessment at the college.

The main variable in this study was the DI of MCQ items. The DI measured the ratio of examinees who responded correctly to an item of all examinees. It was calculated as follows:

$$\text{DI} = \frac{\text{Number of students who responded correctly to an item}}{\text{Total number of students who attempted the same}}$$

The mean DI of each version of the same exam was compared to ensure fairness. Each exam version was then divided into five quintuples, according to the items' order (1–20, 21–40, 41–60, 61–80 and 81–100). Then, similar quintuples from different exams were compiled, and the mean DI of each quintuple was compared using ANOVA and Tukey's tests. A p value of <0.05 was considered statistically significant. The performances of students from different levels throughout those quintuples were also compared.

RESULTS

Five of six exams had three versions, and one exam (C) had only two versions. Table 1illustrates the number of versions for each test and the DI for different versions of the same exam. These findings show no statistically significant difference in the DI between the different versions of the same exam.

As shown in Table 2 and Figure 1, the results revealed no statistically significant differences in the DI of the first three quintuples. By contrast, a significant drop (p <0.05) was observed in the DI of the fourth quintuple, as compared to that of the second and third quintuples. Further remarkable diminution in the DI was observed in the fifth quintuple, as compared to that of the second and third quintuples (p <0.01 and p <0.01, respectively) (Table 3).

Regarding the performance of students according to their level, the study included 514 examinees: 220 freshmen, 178 sophomores and 116 middlers (third-year students). Figure 2 shows that student performance was significantly lowered in the fourth quintuple for all students in their first, second and third years as compared to their performance in the second and third quintuples (p <0.05). In addition, the DI significantly diminished in the fifth quintuple for students in their first and second years (p < 0.01 and p < 0.05, respectively) as compared to that of the third-year students.

DISCUSSION

The effective and reliable measurement of knowledge is an important component of medical education. Educators presume that a student's performance on a test is an indicator of how well the student comprehends the learning materials. A student's performance on a test may be affected by the effort exerted towards learning the material, the student's innate learning ability, random chance and perhaps factors that are related to the test itself **(20, 21)**.

The question of students' performance being affected by the item order in MCQ tests has been explored previously **(22–24)**. The present study found no statistically significant difference in the DI between the different versions of the same test. These findings match those of previous studies, which found that scrambling the item order of MCQs did not affect students' performance **(25, 26)**. However, some previous studies have found that scrambling the item order of MCQs on tests had a statistically significant effect on test scores **(27, 28)**. It is assumed that the level of difficulty of an MCQ test is determined by the level of difficulty of the questions being asked in the test and not the order in which the questions are asked **(29, 30)**. The present study did not align with this assumption.

The findings of the present study showed that the performance of students on a 100-item MCQ test varied remarkably, and their performance increased by the second and third quintuple yet decreased by the fourth quintuple. The increase from the first to the second and third quintuples might be due to anxiety at the beginning of the test, which tends to subside as students become comfortable with the questions. This phenomenon has been reported previously **(31, 32)**. The diminution in the mean DI that was observed in the fourth and fifth quintuples (61–100 items) may be attributed to the exhaustion of students and their suboptimal English language skills, which may have accelerated the loss of their attention and concentration as the test progressed. It seems that the students' performance initially rose as the test progressed, reaching its maximum near the middle, and considerably descended thereafter with time **(26)**.

The number of items that are needed to cover the content and ensure candidate separation reliability is often questioned. On many assessments, reliability has been shown to improve with larger numbers of items on a test **(16)**. Previous studies have reported that two factors can affect the ability of a test to discriminate between levels of student ability: **(1)** the quality of individual test items and **(2)** the number of test items **(4, 7, 16)**.

The present results revealed that item position on MCQ tests significantly affects the performance of students. Variance in student performance supports the statement that students may lose their attention and concentration as tests progress **(31)**. Some authors believe that longer MCQ tests tend to be more reliable because having more items automatically reduces the measurement error **(4, 14, 33)**. However, as MCQ tests become longer, student anxiety increases, while concentration and performance decrease **(31)**. As shown in Figure 1, without considering the quality of test items, increasing the number of questions to increase reliability; is not always beneficial. Instead, it may have a negative impact on student performance.

The current study also addressed whether a student's school level affects the DI of items and therefore his/her performance. The results illustrate that the performance pattern of third-year students was remarkably higher in the last quintuple of the MCQ test compared to that of first- and second-year students. It seems that students who were more advanced in the programme had more testing experience and better English, which enabled them to maintain concentration, even during the last quintuple of the test **(26)**.

The current research presented some limitations, including that it was a one-centre study and it only evaluated students who were in the early years of the programme. Moreover, the number of tests that were examined may not have been sufficient to generalise the findings, and only the DI was evaluated. Therefore, further studies are required using in-depth evaluation of more psychometric parameters and different study designs with larger samples from multiple centres.

Declaration of interest: The author reports no declaration of interest.

Funding: None

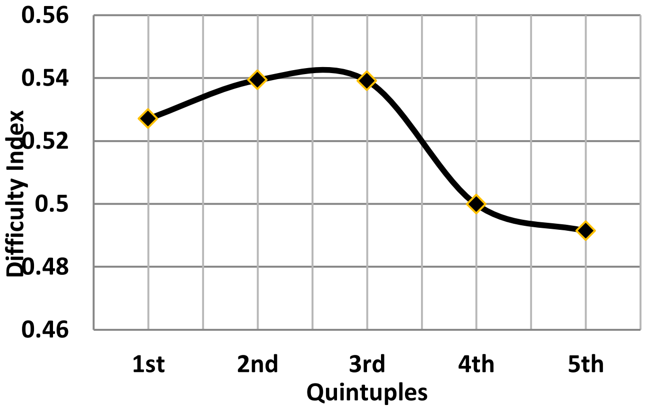

Figure 1. Mean of DI in different quintuples of MCQ exam

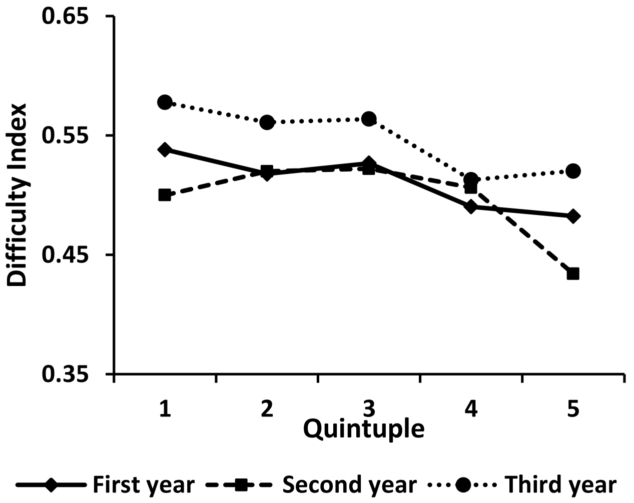

Figure 2. Mean of DI in different quintuples of MCQ exam for the students of first, second and third year

REFERENCES

1. Rudolph MJ, Daugherty KK, Elizabeth RM, Shuford VP, Lebovitz L, DiVall MV. Best practices on exam item construction and post-hoc review. American Journal of Pharmaceutical Education. 2019; ajpe7204 (e-view).
2. Thomas A, Gruppen DL, van der Vleuten C, Chilingaryan G, Amari F, Steinert Y. Use of evidence in health professions education: Attitudes, practices, barriers and supports. Medical teacher. 2019; (in press)
3. Downing SM. Validity: on the meaningful interpretation of assessment data. Medical education. 2003;37(9):830-7.
4. Schuwirth LW, Van Der Vleuten CP. A plea for new psychometric models in educational assessment. Medical education. 2006;40(4):296-300.
5. Sireci SG. Validity issues in accommodating reading tests. Journal of Educators & Education/Jurnal Pendidik dan Pendidikan. 2008;23(7):81-110.
6. Schuwirth LW, Van Der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Medical education. 2004;38(9):974-9.
7. Raymond MR, Grande JP. A practical guide to test blueprinting. Medical teacher. 2019;(in press).
8. Downing SM, Haladyna TM. Validity and its threats. Assessment in health professions education. 2009;1:21-56.
9. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Applied measurement in education. 2002;15(3):309-33.
10. Kehoe J. Basic item analysis for multiple-choice tests. Practical assessment, research & evaluation. 1995;4(10):20-4.
11. Osterlind SJ. Test item bias, quantitative application in the social science: Sage; 1983.
12. Lange A, Lehmann IJ, Mehrens WA. Using item analysis to improve tests. Journal of Educational Measurement. 1967;4(2):65-8.
13. Xing D, Hambleton RK. Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. Educational and Psychological Measurement. 2004;64(1):5-21.
14. Van der Vleuten C. Validity of final examinations in undergraduate medical training. BMJ. 2000;321(7270):1217-9.
15. Schuwirth LW, Van Der Vleuten CP. Current Assessment in Medical Education: Programmatic Assessment. Journal of Applied Testing Technology. 2019; 0(S2):2-10.
16. Downing SM. Reliability: on the reproducibility of assessment data. Medical education. 2004;38(9):1006-12.
17. Josien L, Broderick B. Cheating in higher education: The case of multi-methods cheaters. Academy of Educational Leadership Journal. 2013;17(3):93.
18. Denny P, Manoharan S, Speidel U, Russello G, Chang A, editors. On the Fairness of Multiple-Variant Multiple-Choice Examinations. Proceedings of the 50th ACM Technical Symposium on Computer Science Education; 2019: ACM.
19. Denyer G, Hancock D, editors. Multiple choice questions to combat plagiarism and encourage conceptual learning. Proceedings of The Australian Conference on Science and Mathematics Education (formerly UniServe Science Conference); 2012.

20. West C, Sadoski M. Do study strategies predict academic performance in medical school? Medical education. 2011;45(7):696-703.

21. Tiemeier AM, Stacy ZA, Burke JM. Using multiple choice questions written at various Bloom's Taxonomy levels to evaluate student performance across a therapeutics sequence. Innovations in pharmacy. 2011;2(2):1-11.

22. Sue DL. The effect of scrambling test questions on student performance in a small class setting. Journal for Economic Educators. 2009;9(1):32-41.

23. Davis DB. Exam Question Sequencing Effects and Context Cues. Teaching of Psychology. 2017;44(3):263-7.

24. Garbarski D, Schaeffer NC, Dykema J. The effects of response option order and question order on self-rated health. Quality of Life Research. 2015;24(6):1443-53.

25. Neely DL, Springston FJ, McCann-Stewart J. Does Item Order Affect Performance on Multiple-Choice Exams? Handbook for Teaching Introductory Psychology: With an Emphasis on Assessment. 2001:108.

26. Perlini AH, Lind DL, Zumbo BD. Context effects on examinations: The effects of time, item order and item difficulty. Canadian Psychology/Psychologie canadienne. 1998;39(4):299.

27. Carlson JL, Ostrosky AL. Item sequence and student performance on multiple-choice exams: Further evidence. The Journal of Economic Education. 1992;23(3):232-5.

28. Doerner WM, Calhoun JP. The Impact of the Order of Test Questions in Introductory Economics. Available at SSRN 1321906. 2009.

29. Anderson J. The MCQ controversy—a review. Medical Teacher. 1981;3(4):150-6.

30. Alsubait T, Parsia B, Sattler U, editors. A similarity-based theory of controlling mcq difficulty. 2013 Second International Conference on e-Learning and e-Technologies in Education (ICEEE); 2013: IEEE.

31. Kiat JE, Ong AR, Ganesan A. The influence of distractor strength and response order on MCQ responding. Educational Psychology. 2018;38(3):368-80.

32. Avcu A, Tunç EB, Uluman M. How the order of the items in a booklet affects item functioning: emprical findings from course level data? European Journal of Education Studies. 2018;4(3):227-239.

33. Meshkani Z, Abadie FH. Multivariate analysis of factors influencing reliability of teacher made tests. Journal of Medical Education. 2005;6(2):149-152.

Table 1. Mean of difficulty index (DI) of the different versions of the MCQ exams

Preclinical Courses	**Year level**	**No of versions**	**Mean DI (± SD)**			***P*-value (Significance)**
			Version 1	**Version 2**	**Version 3**	
Exam A	First year	3	0.53 (±0.24)	0.49 (±0.23)	0.50 (±0.26)	> 0.05 (NS)
Exam B	First year	3	0.61 (±0.24)	0.61 (±0.24)	0.62 (±0.27)	> 0.05 (NS)
Exam C	Second year	2	0.54 (±0.21)	0.57 (±0.23)	NA	> 0.05 (NS)
Exam D	Second year	3	0.46 (±0.20)	0.45 (±0.21)	0.46 (±0.21)	> 0.05 (NS)
Exam E	Third year	3	0.56 (±0.22)	0.53 (±0.21)	0.54 (±0.21)	> 0.05 (NS)
Exam F	Third year	3	0.63 (±0.22)	0.62 (±0.21)	0.65 (±0.20)	> 0.05 (NS)

Table 2. Descriptive statistical analysis of item difficulty index in different quintuples

Quintuple	Item order	Mean (DI)	Variance	SD	SE
1st	**(1-20)**	0.5265	0.06172	0.248	0.00778
2nd	**(21-40)**	0.5388	0.05642	0.237	0.00743
3rd	**(41-60)**	0.5385	0.06119	0.247	0.00774
4th	**(61-80)**	0.4993	0.06235	0.249	0.00782
5th	**(81-100)**	0.4908	0.05821	0.241	0.00755

Table 3. Tukey's test of significance between different quintuples of exams

Quintuples	1st	2nd	3rd	4th	5th
1st		N.S.	N.S.	N.S.	<0.05
2nd	N.S.		N.S.	<0.05	<0.01
3rd	N.S.	N.S.		<0.05	<0.01
4th	N.S.	<0.05	<0.05		N.S.
5th	<0.05	<0.01	<0.01	N.S.	

N.S refers to Non-staticsally significance

How to cite this article: Alamro A. S., The Effect Of Order Of Mcq Items On Difficulty Index?. Int.J.Med.Sci.Educ 2019; 6(2): 1-6