SHORT COMMUNICATION

Volume 16 Supp.1 2024

DOI: 10.21315/eimj2024.16.s1.3

ARTICLE INFO

Received: 02-03-2024
Accepted: 31-03-2024
Online: 31-07-2024

Enhancing Assessment towards Flexible Medical Education: A Deep Dive into Preclinical Short Answer Questions through Item Analysis

Azli Shahril Othman[1], Sara Idris[2], Atikah Abdul Latiff[2], Norhafizah Ab Manan[2], Norfaizatul Shalida Omar[3]

[1]*Faculty of Medicine and Defence Health, Universiti Pertahanan Nasional Malaysia, Kuala Lumpur, MALAYSIA*

[2]*Faculty of Medicine, University of Cyberjaya, Selangor, MALAYSIA*

[3]*Faculty of Medicine, Manipal University College Malaysia, Melaka, MALAYSIA*

To cite this article: Othman AS, Idris S, Abdul Latiff A, Ab Manan N, Omar NS. Enhancing assessment towards flexible medical education: a deep dive into preclinical short answer questions through item analysis. Education in Medicine Journal. 2024;16(Supp.1):23–8. https://doi.org/10.21315/eimj2024.16.s1.3

To link to this article: https://doi.org/10.21315/eimj2024.16.s1.3

ABSTRACT

Short answer questions (SAQs) and other forms of similarly structured examination methods are employed as evaluation tools to gauge the competency of medical students at various levels of study. However, item analyses for these questions have rarely been conducted. Evaluating the quality of examination questions through item analysis ensures that stakeholders, especially those undergoing learning in various flexible pathways in medical education, are provided with reliable and relevant assessments, promoting effective learning and competency development. In this study, we performed item analyses on SAQs by extracting the passing index (PI) and discrimination index (DI) for each sub-question. Data were analysed using Microsoft Excel (Microsoft Corporation, United States) and Jeffreys's Amazing Statistics Program (JASP) (University of Amsterdam, The Netherlands). Twenty-seven sub-questions from five SAQs were analysed. The DI of the sub-questions ranged from 0.043 to 0.935 with a mean of 0.449 ± 0.223, while the PI returned a range of 0.012 to 0.971 with a mean value of 0.597 ± 0.246. In conclusion, the SAQs administered during a professional examination of preclinical medical students exhibited an acceptable range and mix of PI and DI values. However, improvements must be made to the sub-questions that return poor PI and DI values.

Keywords: *Discrimination index, Item analysis, Short answer questions, Assessment*

CORRESPONDING AUTHOR

Azli Shahril Othman, Faculty of Medicine and Defence Health, Universiti Pertahanan Nasional Malaysia, 57000 Kuala Lumpur, Malaysia

Email: azli.shahril@upnm.edu.my

INTRODUCTION

Assessment remains a fundamental tool for the training of medical students. Assessments allow educators to gauge the knowledge and skills acquired by students and provide a comprehensive picture of their progress throughout training. Even though lecturers continue to lean more towards entrustable, formative, and reflective strategies of learning and assessment, summative examination and evaluation methods such as multiple-choice questions (MCQs), one-best answer questions, short answer questions (SAQs), and modified

essay questions are the main choices used to gauge the continuous progress and overall academic standing of medical students (1). These evaluation approaches offer robust and reliable methods to measure the performance of medical students. With artificial intelligence and online learning fuelling flexible learning pathways for students to complete their undergraduate medical studies (2), it is crucial that the quality of these evaluation methods, particularly the questions which compose the essence of these assessments, are subjected to rigorous and continuous quality assurance tests to consistently improve their standards.

Item analysis is a statistical technique used to evaluate the quality and effectiveness of individual test items or questions during an assessment (3). It involves analysing the performance data of students on each item to assess difficulty, discrimination (ability to discriminate between high-performing and low-performing test-takers), and overall effectiveness. The passing index (PI), or difficulty index, and discrimination index (DI) are the two most widely reported statistics of item analysis discussed in the medical education literature. Other parameters such as reliability and distractor efficiency are often deliberated upon in item analyses of MCQs. The results obtained from item analysis provide valuable insights into the quality of the assessment, help identify problematic items, and guide the improvement of future assessments. Item analysis allows educators to determine whether the items used for testing are appropriately challenging and aligned with learning objectives. Analysing the difficulty level of items helps ensure that the assessment is neither too easy nor too difficult, providing a valid and reliable measure of students' knowledge and skills. Additionally, identifying items with poor discrimination allows educators to revise, eliminate, and improve the assessment's ability to differentiate student performance accurately (4).

Item analysis of MCQs is regularly reported and discussed in the literature. In a study examining 1,500 MCQs for a medical licencing examination in Mongolia, the number of questions with a zero or negative DI was as high as 11.6%, whereas questions which were deemed to be of high PI were 21.9% (5). Another team of researchers looking to optimise a question bank for MCQs in ophthalmology found that the mean PI and mean DI for their set of questions were 0.532 and 0.260, respectively (6). Based on the values reported in the literature, we can ensure that medical educationists, content experts, and faculty members continuously improve the quality and effectiveness of these questions.

In contrast, item analyses of SAQs and other forms of similarly structured examination methods are rarely reported in the medical education literature. This may be due to the limited availability and functionality of automated item analysis methods for SAQs (7), unlike the software applications available for the analysis of MCQs. Therefore, we sought to determine the quality and effectiveness of these SAQs by examining the PI of the item analysis. We also examined how these questions can differentiate between students who performed relatively well and those who performed relatively worse on the examination by investigating the DI.

METHODS

A total of 169 second-year medical students who participated in a preclinical professional examination at an institution of higher learning in Cyberjaya, Selangor, Malaysia, were sorted according to their total marks obtained in the SAQs component of the examination. Five SAQs were administered and further categorised into 27 sub-questions. The students were given 1 hour and 40 minutes to answer all questions, and the marks obtained comprised

30% of their final professional examination score. A sub-question was deemed correct if a student obtained at least 50% of the total marks allocated to that sub-question and incorrect if a student obtained less than 50% of the total marks. The PI for each sub-question was then generated according to the percentage of students who answered the question correctly.

For the DI, data from the top 27% (n = 46) and bottom 27% (n = 46) group of students who attempted the examination were isolated for the analysis. The DI was then determined using the following formula: [(number of students with correct answers in the top 27% group – number of students with correct answers in the bottom 27% group)/total number of students in one group].

All 27 sub-questions were classified according to Bloom's taxonomy of cognitive domain levels (8). This classification process was led by each primary question author with support and feedback from content experts. Content experts analysed each sub-question separately, determined the level of cognitive complexity based on the key elements, action verbs, and sentence structure, and assigned an appropriate level to the question.

All data were analysed using Jeffreys's Amazing Statistics Program (JASP) (University of Amsterdam, The Netherlands) and Microsoft Excel (Microsoft Corporation, United States). Relevant data were presented as mean ± standard deviation (SD), and Pearson's correlation coefficient (r) was computed to assess the relationship between the PI and DI. Statistical significance was set at $p < 0.05$.

RESULTS

The PI of the 27 sub-questions ranged from 0.012 to 0.971 with a mean value of 0.597 ± 0.246. The DI of the 27 sub-questions ranged from 0.043 to 0.935, with a mean value of 0.449 ± 0.223. Of the 27 sub-questions, three sub-questions had relatively low DI values of 0.043, 0.152, and 0.130, which corresponded to PI values of 0.971, 0.941, and 0.935, respectively. None of the sub-questions returned a negative DI. The DI negatively correlated with the PI, with a Pearson's r value of –0.463 ($p < 0.05$). The PI and DI values, together with Bloom's taxonomy cognitive domain level for each sub-question, are given in Table 1.

Table 1: PI, DI and Bloom's taxonomy cognitive domain levels of each sub-question

Question number	PI	DI	Bloom's taxonomy cognitive domain levels
1	0.476	0.543	Comprehension
2	0.441	0.674	Comprehension
3	0.547	0.478	Comprehension
4	0.453	0.609	Comprehension
5	0.647	0.739	Application
6	0.971	0.043	Knowledge
7	0.829	0.239	Comprehension
8	0.788	0.196	Knowledge
9	0.941	0.152	Comprehension
10	0.518	0.326	Knowledge
11	0.653	0.652	Comprehension
12	0.447	0.500	Knowledge

(*Continued on next page*)

Table 1: (*Continued*)

Question number	PI	DI	Bloom's taxonomy cognitive domain levels
13	0.712	0.239	Knowledge
14	0.312	0.457	Comprehension
15	0.382	0.522	Comprehension
16	0.853	0.283	Application
17	0.912	0.217	Knowledge
18	0.782	0.630	Comprehension
19	0.012	0.239	Comprehension
20	0.424	0.674	Knowledge
21	0.894	0.500	Knowledge
22	0.935	0.130	Application
23	0.406	0.543	Comprehension
24	0.359	0.935	Comprehension
25	0.318	0.609	Comprehension
26	0.700	0.674	Comprehension
27	0.412	0.326	Knowledge

DISCUSSION

Our study showed that the SAQs deployed in this preclinical professional examination were generally of good quality. There was a good range of easy, moderate, and difficult questions as shown by the PI range of 0.012 to 0.971 and mean value of 0.597 ± 0.246. Our findings are consistent with those of a recent study which focused on the discrimination power of short essay questions versus MCQs in the evaluation of preclinical medical students (7). Their study, which extracted data from 34 short essay questions attempted by 726 students, returned PI and discrimination factor values of 0.73 ± 0.03 and 0.68 ± 0.01, respectively. However, their study focused only on questions involving preclinical biochemistry topics, whereas our questions included a wide range of questions from various preclinical disciplines such as anatomy, biochemistry, physiology, microbiology, pathology, and pharmacology.

The questions analysed in our study exhibited a DI range of 0.043–0.935, with a mean value of 0.449 ± 0.223. Three sub-questions with the lowest DI values returned high corresponding PI values; this has been shown to be similar in item analyses of other types of examination questions, such as MCQs (9, 10). Our results are further supported by the negative correlation between DI and PI (Pearson's $r = -0.463$ [$p < 0.05$]); easier questions do not discriminate well between higher-performing and lower-performing students. However, in another study examining a set of 257 MCQs administered to final-year medical students over five years, it was found that factual questions with lower PI values returned higher discriminative index scores in students with moderate or poor academic performance than in students with good academic performance (11).

Notably, in our study, the three questions with relatively poor DI and high PI did not correspond to a particular category under Bloom's taxonomy of cognitive domain levels. Of the seven questions with a high PI of > 0.8, three were knowledge-, two were comprehension-, and two were application-type questions. Questions with poor or marginal DI of < 0.29 were also classified according to Bloom's taxonomy, with three knowledge-, three comprehension-, and two application-type questions. Previously, it was shown that

the DI values for application and synthesis or evaluation type of questions were significantly higher than those for knowledge and comprehension type questions (12). In their study, the types of questions analysed were MCQs, and a much larger pool of questions to perform their analysis. They also found only a 54% match between the expected difficulty level (as determined by experts) and actual difficulty level (as answered by students in the examination).

The findings of our study reiterate the importance of performing item analysis on SAQs and similar examination methods such as modified essay questions. While the process of analysing these questions may be tedious and time-consuming, the findings provide insight into the actual quality of the questions, which may have been overlooked and/or under-analysed in the past. Questions that are relatively easy for students but do not sufficiently discriminate between higher-performing and lower-performing students should be further scrutinised, and the value of these questions as part of a summative assessment of students should be determined. With the advent of examination technology and increasing availability of automated item analysis software, quick and convenient item analysis is no longer limited to MCQs but can be expanded to other types of examination questions. Consistent quality assurance ensures that examination questions remain of high quality, align with the intended learning outcomes, meet established standards set by the teaching institution, and enable continuous enhancement of assessment and evaluation methods for flexible learning pathways in medical education.

CONCLUSION

SAQs administered during professional examinations of preclinical medical students were within an acceptable range of difficulty and exhibited sufficient discriminating power. However, improvements should be made to the sub-questions that return poor PI and DI values. Item analyses of SAQs should also be performed regularly to gauge the overall quality of the test items.

ACKNOWLEDGEMENTS

We would like to thank the lecturers from the Division of Basic Medical Sciences, Faculty of Medicine, University of Cyberjaya for their effort and expertise in acting as question authors and examiners for the examination from which the data for this study were taken.

REFERENCES

1. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med. 2019;94(6):902–12. https://doi.org/10.1097/ACM.0000000000002618

2. Shankar PR, Azhar T, Nadarajah VD, Er HM, Arooj M, Wilson IG. Faculty perceptions regarding an individually tailored, flexible length, outcomes-based curriculum for undergraduate medical students. Korean J Med Educ. 2023;35(3):235–47. https://doi.org/10.3946/kjme.2023.262.

3. McCrossan P, Nicholson A, McCallion N. Minimum accepted competency examination: test item analysis. BMC Med Educ. 2022;22(1):400. https://doi.org/10.1186/s12909-022-03475-8

4. Owolabi LF, Adamu B, Taura MG, Isa AI, Jibo AM, Abdul-Razek R, et al. Impact of a longitudinal faculty development program on the quality of multiple-choice question item writing in medical education. Ann Afr Med. 2021;20(1):46–51. https://doi.org/10.4103/aam.aam_14_20

5. Gomboo A, Gombo B, Munkhgerel T, Nyamjav S, Badamdorj O. Item analysis of multiple-choice questions in medical licensing examination. Cent Asian J Med Sci. 2019;5(2):141–8. https://doi.org/10.24079/CAJMS.2019.06.009

6. Bhat SK, Prasad KHL. Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology: a cross-sectional study. Indian J Ophthalmol. 2021;69(2):343–6. https://doi.org/10.4103/ijo.IJO_1610_20

7. Eldakhakhny B, Elsamanoudy AZ. Discrimination power of short essay questions versus multiple choice questions as an assessment tool in clinical biochemistry. Cureus. 2023;15(2):e35427. https://doi.org/10.7759/cureus.35427

8. Mohammed M, Omar N. Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. PLOS ONE. 2020;15(3):e0230442. https://doi.org/10.1371/journal.pone.0230442

9. Bhattacherjee S, Mukherjee A, Bhandari K, Rout AJ. Evaluation of multiple-choice questions by item analysis, from an online internal assessment of 6th semester medical students in a Rural Medical College, West Bengal. Indian J Community Med. 2022;47(1):92–5. https://doi.org/10.4103/ijcm.ijcm_1156_21

10. Kumar D, Jaipurkar R, Shekhar A, Sikri G, Srinivas V. Item analysis of multiple-choice questions: a quality assurance test for an assessment tool. Med J Armed Forces India. 2021;77(Suppl 1):S85–9. https://doi.org/10.1016/j.mjafi.2020.11.007

11. Iñarrairaegui M, Fernández-Ros N, Lucena F, Landecho MF, García N, Quiroga J, et al. Evaluation of the quality of multiple-choice questions according to the students' academic level. BMC Med Educ. 2022;22(1):779. https://doi.org/10.1186/s12909-022-03844-3

12. Hamamoto Filho PT, Silva E, Ribeiro ZMT, Hafner MLMB, Cecilio-Fernandes D, Bicudo AM. Relationships between Bloom's taxonomy, judges' estimation of item difficulty and psychometric properties of items from a progress test: a prospective observational study. Sao Paulo Med J. 2020;138(1):33–9. https://doi.org/10.1590/1516-3180.2019.0459.R1.19112019