ORIGINAL ARTICLE
Volume 11 Issue 4 2019
DOI: 10.21315/eimj2019.11.4.3
ARTICLE INFO
Submitted: 08-10-2019
Accepted: 13-12-2019
Online: 31-12-2019

An Analysis of the Student's Performance and Marking Process in Open-Ended Short Answer Questions in a Medical Programme: A Pilot Study

Thin Thin Win[1], Katrina Chung Pooi Yin[1], Sunil Pazhayanur Venkateswaran[1], Shadi Khadijeh Gholami[2], Nyunt Wai[2]

[1]*Division of Pathology, School of Medicine, International Medical University, Kuala Lumpur, MALAYSIA*

[2]*Division of Human Biology, School of Medicine, International Medical University, Kuala Lumpur, MALAYSIA*

To cite this article: Win TT, Yin KCP, Venkateswaran SP, Gholami SK, Wai N. An analysis of the student's performance and marking process in open-ended short answer questions in a medical programme: a pilot study. Education in Medicine Journal. 2019;11(4):29–35. https://doi.org/10.21315/eimj2019.11.4.3

To link to this article: https://doi.org/10.21315/eimj2019.11.4.3

ABSTRACT

There has been no literature on analysing short answer questions (SAQs) to assess how well the students captured the essence of core concepts in basic sciences disciplines. The objective of this study was to assess the performance of students and marking process in core concept-intensive SAQs on physiology and pathology in a professional examination of the undergraduate medical curriculum. Two faculties, each from physiology and pathology disciplines identified the core concept–intensive SAQs. Consensus was taken on how core are the concepts and a marking scheme was developed by ranking of the elements of the concept. Based on their overall performance, top 20 students, middle 20 students and bottom 20 were identified, and their answers were graded (parameters: not attempted, wrong concept, incomplete or irrelevant, and how well the essence of critical key concepts was captured). This communication focused on the discrepancies in the marks awarded by the marker (examiner) and by the investigators. The discrepancy was higher in the middle 20 group while most of the bottom 20 had no marks awarded as they had not attempted or had a wrong or irrelevant concept. The construction of model answers in the examinations should be improved to unlock how much students could capture the underlying key critical concepts. Alternatively, the marker should be a content expert capable of going beyond the given model answer.

Keywords: *Core concept, Open-ended short answer questions, Undergraduate medical curriculum*

CORRESPONDING AUTHOR

Dr. Thin Thin Win, Division of Pathology, School of Medicine, International Medical University, 126, Jalan Jalil Perkasa 19, Bukit Jalil, 57000 Kuala Lumpur, Malaysia | Email: thinthinwin@imu.edu.my

INTRODUCTION

In the assessment of a medical programme, various written assessment tools have been used to assess the understanding of knowledge and concepts of medical students. One of these tools which is widely used is the short answer questions (SAQs), an open-ended and semi-structured question format which possesses higher reliability and validity (1). Although the higher order learning can be assessed by multiple choice questions (MCQs) such as single best answer type, multiple true

false types and script concordance test, through application of basic concepts; open-ended SAQs can directly unlock how much students master these concepts by letting them express in writing. Students are required to generate an answer spontaneously in response to SAQs (2) and it possesses higher reliability and validity with less cueing, shows higher discrimination (1).

MCQs may not provide a true reflection of knowledge as they rely on answer recognition rather than recall (3–4). Some MCQs formats could be limited to lower order such as one true answer rather than one best answer. However, SAQs can assess student's ability to synthesise or generate rather than to recognise a correct answer, it provides greater validity (5). SAQs are commonly used for the evaluation of higher-order cognitive skills including interpretation, problem-solving and reasoning; and it minimises the cueing effect. Reliability can be increased with the availability of clear grading guidelines and training of markers. Essays are used for evaluation of more complex cognitive processes including interpretation, problem solving and synthesis of information. The testing time is relatively long and the grading can be time consuming. A comprehensive marking scheme is required to increase inter-rater reliability (1, 6).

There is an extensive literature on post-examination analysis (psychometrics) for MCQs (7) as well as essay questions (8); but not on SAQs to assess how well students captured the essence of key concepts in basic sciences disciplines. Some studies on SAQs reported that mean standard deviation of SAQs score is wider than that of MCQs (9). Hence, this study was carried out to fill this gap as it is becoming evident that with knowledge explosion, learning should be by understanding and application of the core or threshold concepts in a discipline and not by content memorisation (10–11).

MATERIALS AND METHODS

This is a retrospective pilot study on the analysis of total 60 students' performance on SAQs in Semester 3, Year 2 written professional examination of the medical programme in the International Medical University. The SAQs paper was blueprinted according to the learning outcomes of the Bachelor of Medicine, Bachelor of Surgery (MBBS) curriculum and covers contents across six modules taught in Semesters 2 and 3. The paper underwent vetting at three levels (module, semester and dean's vetting) and also standard set using the modified Angoff method. The question authors have been trained through various faculty development activities on how to construct good quality SAQs. Two faculties each from the physiology and pathology disciplines identified the core concept–intensive SAQs in the two disciplines (physiology and pathology). A total of nine SAQs which covered core concepts of physiology and pathology disciplines were collected from the Semester 3, Year 2 written professional examination of the medicine programme (Table 1). Consensus between the two faculties was taken on the core concepts tested and a marking scheme was developed by ranking of the elements of the core concept as illustrated in Figures 1 and 2.

Based on the cohort's overall performance in the written components of the first professional examination at the end of Semester 3, 60 students were identified and divided into the top 20 students, middle 20 students and bottom 20, and their answers were graded. The parameters used for the grading of the answers were not attempted, wrong concept, incomplete or irrelevant, and how well the essence of critical key concepts was captured. The students' answers and marks given by the original markers were reviewed.

Table 1: Nine SAQs which covered core concepts of physiology and pathology disciplines

SAQs no.	Area covered	Disciplines involved
1	Heart failure	Physiology, Pathology*
2	Hypertension	Physiology, Pathology*
3	Bronchial asthma	Pathology*
4	Erythrocyte sedimentation	Physiology, Pathology*
5	Acute myeloid leukaemia	Pathology*
6	Emphysema	Physiology, Pathology*
7	Glucose metabolism	Physiology*
8	Sympathetic activity on myocardial blood flow	Physiology*
9	Valvular heart disease	Physiology, Pathology*

Note: *Refers to the discipline who primarily set the question.

Question:

Describe the haemodynamic effects of stenosis of aortic valve. (2 marks)

Answer:

Stenosis is the **failure of a valve to open completely (1+)**, **obstructing forward flow (1+)** and reducing the amount of blood ejected by the **left ventricle** into the aorta (**reduced stroke volume**) (4+). Hence the **end-systolic volume and pressure** in the **left ventricle increases** (4+).

Grading of core: Most critical core element (4+), Lesser critical core elements (3+) (2+) and Least critical core element (1+)

Figure 1: Example of ranking of core elements (Physiology).

Question:

Define Emphysema. (2 marks)

Answer:

Emphysema is a condition of the lung characterised by **irregular irreversible enlargement of the airspaces distal to the terminal bronchioles (4+)** accompanied by **destruction of their walls (2+) without obvious fibrosis (1+).**

Grading of Core: Most critical core element (4+), Lesser critical core elements (3+) (2+) and Least critical core element (1+)

Figure 2: Example of ranking of core elements (Pathology).

The students' performances were graded into five categories: 0 (not attempted), 0 (attempted), minimum, moderate and full. Category 0 (not attempted) was given to those who did not attempt the question. Category 0 (attempted) was given to those who attempted with a totally wrong answer and concept. Minimum category was given to those who scored less than half of the total mark. Moderate category was given to those who scored half of the total mark. Full category was given to those who scored the full mark for the question. This analysis focused on the discrepancies in the marks awarded by the original marker (examiner) and by the investigators. In the event of any discrepancies, consensus was obtained between two investigators of the same discipline.

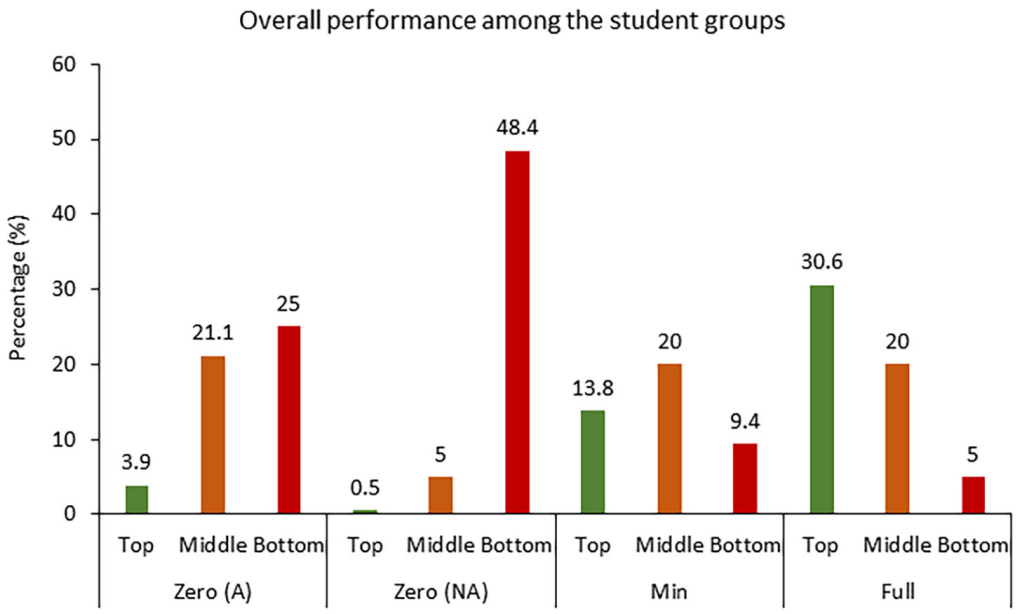
RESULTS

A total of nine SAQs which covered core concepts of physiology and pathology discipline from the Semester 3, Year 2 written professional examination of the medicine programme, were included in this study. The overall performance of the top, middle and bottom groups on the nine SAQs were analysed and results were as shown in Table 2 and Figure 3. The discrepancies and total agreement percentages between original markers and investigators are shown in Table 3.

Table 2: Overall performance of the top, middle and bottom group

No.	Top 20					Middle 20					Bottom 20				
	Zero		Min	Mod	Full	Zero		Min	Mod	Full	Zero		Min	Mod	Full
	A	NA				A	NA				A	NA			
1	1	0	0	10	9	0	0	0	10	10	0	13	0	3	4
2	1	0	2	13	4	6	3	2	9	0	0	20	0	0	0
3	4	0	1	11	5	8	8	10	8	2	12	3	2	3	0
4	0	0	0	3	17	0	0	0	4	16	1	3	0	11	5
5	0	1	1	9	8	11	4	3	0	2	10	8	0	2	0
6	0	0	12	6	2	5	1	9	5	0	10	6	4	0	0
7	0	0	5	15	0	2	0	10	8	0	6	6	8	0	0
8	1	0	3	12	4	6	1	0	9	4	6	11	0	3	0
9	0	0	1	13	6	8	0	2	8	2	0	17	3	0	0
Total	7	1	25	92	55	38	9	36	61	36	45	87	17	22	9
%	3.9	0.5	13.8	51.2	30.6	21.1	5	20	33.9	20	25	48.4	9.4	12.2	5

Note: A: attempt; NA: not attempt; Min: minimum; Mod: moderate.



Note: A: attempt; NA: not attempt; Min: minimum; Mod: moderate.

Figure 3: Bar charts showing the performance of students in each of the three groups.

Table 3: Discrepancy and total agreement percentage in marks awarded between markers and investigators

SAQs number	Discrepancy (number of students)			Total agreement %
	Bottom 20	Middle 20	Top 20	
1	0	2	2	93.3
2	0	0	1	98.3
3	0	4	1	91.7
4	1	6	3	83.3
5	3	3	2	86.7
6	9	9	13	48.3
7	0	1	1	96.7
8	0	0	1	98.3
9	0	5	3	86.7

Students who obtained zero were noted that they either did not attempt the questions at all or attempted but with irrelevant or wrong concepts. Almost half (48.4%) from the bottom group did not attempt the questions, compared with 0.5% and 5% from the top and middle groups, respectively. For answers with fully acceptable concepts as shown in Figures 1 and 2, 30.6% came from the top group, 20% from the middle and 5% from the bottom group (Table 2 and Figure 3).

In the top 20 group, 3.9% have a totally wrong concept and 13.8% gave answers which are minimally acceptable. The respective findings for the middle group are 21.1% and 20%, and 25% and 9.4% for the bottom group. The discrepancy between markers and investigators was higher in the middle group while most of the bottom group had no marks awarded as they had not attempted and had a wrong or irrelevant concept.

DISCUSSION

In this study, according to the analysis of the performance of students answering SAQs and the marking process with model answers, the bottom group students could not grasp the concept of the questions; hence, they did not attempt the question at all or attempted the question with totally wrong concepts. On the other hand, 5% and 12.2% from the bottom group can score full marks and moderate score respectively, which could be due to correctly spotting the topics for the examination or easy questions or question with high difficult index or presence of cues in some questions.

In this study, top 20 and middle 20 group scored moderate and full marks. However, this observation might not fully reflect their understanding as it can be confounded by construct irrelevant variances such as guessing, flair of writing. Students in the top and middle 20 groups may have wrong concepts, due to poor understanding or bluffing as there is no negative marking for the wrong answers. It can be inferred that not all students from the top group had the grasp of correct concepts. In the failure of scoring marks, spotting questions is likely to be the cause in the case of bottom group while poor time management might be the cause for the top group.

A study revealed that an open-ended format is not inherently better at assessing higher order cognitive skills than MCQs (2). The MCQs can test higher order skills as effectively as the SAQs and can be used as a single format in written assessment provided

quality items testing higher order cognitive skills (12). However, when properly created, SAQs is a reliable, constructed-response assessment tool. It can be consistently and fairly scored. Students were asked to write down their answers to a given question instead of selecting from a list of provided options like multiple-choice and matching formats (13). However, students' performance can be different between MCQs and SAQs. A study showed that more students had scores greater than the median in the SAQs scores (14). Another study found that a statistically significant overall correlation between students' performance on MCQs and SAQs in major clinical subjects (9).

We found that there were discrepancies between the markers and investigators. It was higher in the middle group while most of the bottom group had no marks awarded as they had not attempted, had a wrong or irrelevant concept. The reverse could be true for the top group as the students from this group were able to answer most of the questions correctly. SAQ no.6 with the largest discrepancy is on the definition of bronchial asthma, where the marker had strictly followed the model answer.

In SAQs marking, there is a wider standard deviation by bias or variable marking by different examiners whereas marking is more consistent in MCQs (9). However, SAQs have high validity as students need to be able to generate the piece of knowledge in the absence of cues, an approach that is more representative of real-life medical practice (5). However, scores awarded in SAQs may be influenced by student factors, marker factors and quality of questions with their model answers. In marking SAQs, markers do not necessarily need to follow the model answer strictly. If the marker is a content expert, marks can be awarded if the student's answer is within the core concept.

If multiple examiners are available, double marking is preferred in marking SAQs for more reliable scores as discrepancy is reduced through cross-checking between examiners. For efficiency, however, each marker should correct the same question for all candidates. This leads to more reliable scores than if each marker corrects all the questions of one group of candidates while another marker corrects all questions for another group (15).

In reality, every question format has its own advantages and disadvantages which must be carefully weighed when a particular question type is chosen. It is not possible that one type of question will cover all the aspects of a topic. Therefore, a variety of formats are needed to counter the possible bias associated with individual formats and they should be consistent with the stated objectives of the course or programme (16).

CONCLUSION

This study revealed that an appreciable number of students including the top group did not demonstrate the grasp of correct core knowledge in answering SAQs. The construction of model answers in the examinations should be improved to unlock how much students could capture the underlying key critical concepts. Alternatively, the marker should be a content expert capable of going beyond the given model answer. As time factor can influence the quality marking process, enough duration for the marking process is important for the markers. As this study is a pilot study, there are some limitation that all SAQs cannot be assessed. For more precised results, larger number of students with larger number of SAQs that cover other disciplines of medical sciences should be included.

REFERENCES

1. Sam AH, Field SM, Collares CF, van der Vleuten CPM, Wass VJ, Melville C, et al. Very-short-answer questions: reliability, discrimination and acceptability. Med Educ. 2018;52(4):447–55. https://doi.org/10.1111/medu.13504

2. Hift RJ. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? BMC Med Educ. 2014;14:249.

3. Elstein AS. Beyond multiple-choice questions and essays: the need for a new way to assess clinical competence. Acad Med. 1993;68(4):244–9. https://doi.org/10.1097/00001888-199304000-00002

4. Veloski JJ, Rabinowitz HK, Robeson MR, Young PR. Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence. Acad Med. 1999;74(5):539–46. https://doi.org/10.1097/00001888-199905000-00022

5. Sam AH, Hameed S, Harris J, Meeran K. Validity of very short answer versus single best answer questions for undergraduate assessment. BMC Med Educ. 2016;16(1):266. https://doi.org/10.1186/s12909-016-0793-z

6. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple-choice questions? Research paper. BMC Med Educ. 2007;7:49. https://doi.org/10.1186/1472-6920-7-49

7. Mohsen TM, Dennick R. Post-examination analysis of objective tests. Med Teach. 2011;33(6):447–58. https://doi.org/10.3109/0142159X.2011.564682

8. Chan WS. Optimal discrimination index and discrimination efficiency for essay questions. J Appl Meas. 2014;15(1):94–9.

9. Farooqui F, Saeed N, Aaraj S, Sami MA, Amir M. A comparison between written assessment methods: multiple-choice and short answer questions in end-of-clerkship examinations for final year medical students. Cureus. 2018;10(12):e3773. https://doi.org/10.7759/cureus.3773

10. Neve H, Wearn A, Collett T. What are threshold concepts and how can they inform medical education? Med Teach. 2016;38(8):850–3. https://doi.org/10.3109/0142159X.2015.1112889

11. Michael J, Modell H, McFarland J, Cliff W. The "core principles" of physiology: what should students understand? Adv Physiol Educ. 2009;33:10–6. https://doi.org/10.1152/advan.90139.2008

12. Pham H, Trigg M, Wu S, O'Connell A, Harry C, Barnard J, et al. Choosing medical assessments: does the multiple-choice question make the grade? Educ Health (Abingdon). 2018;31(2):65–71. https://doi.org/10.4103/efh.EfH_229_17

13. Royal College of Physician and Surgeon of Canada. Short-answer questions: guidelines for their development. Editorial revisions. Canada: Royal College of Physician and Surgeon of Canada; 2007 [cited 1 October 2019]. Available from: https://saineville.com/wp-content/uploads/2016/11/RC-SAQ-Guidelines.pdf

14. Adamu LH, Yusuf AH, Taura MG. Differences in multiple choice and short answer questions performance among medical and dental trainees as reflected in their anatomy course. J Contemp Med Educ. 2018;6(1):7–16. https://doi.org/10.5455/jcme.20170706060134

15. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ. 2004;38:974–9. https://doi.org/10.1111/j.1365-2929.2004.01916.x

16. Al-Wardy NM. Assessment methods in undergraduate medical education. Sultan Qaboos Univ Med J. 2010;10(2):203–9.