Bull. Egypt. Soc. Physiol. Sci. Vol. (44), Issue (1), 29-38

Bull. of Egyp. Soc. Physiol. Sci.

(Official Journal of Egyptian Society for Physiological Sciences)

(pISSN: 1110-0842; eISSN: 2356-9514)

Multiple choice questions as a tool for summative assessment in medical schools

Saeed Awad M. Alqahtani

Associate professor of Physiology, Department of Physiology, College of Medicine, Taibah University, Medina, Saudi Arabia

Submit Date : 22 Sept. 2023
Revise Date : 31 Oct. 2023
Accept Date: 09 Dec. 2023

Keywords

- Assessment
- Multiple choice questions
- Reliability
- Validity
- Item analysis

Abstract

Objectives: To evaluate the quality of multiple-choice questions (MCQs) used in a summative assessment of a Central Nervous System (CNS) module at the Faculty of Medicine, Jazan University. Methods: Item analysis was conducted on a 70-item MCQ exam administered to 57 medical students after completing the CNS module. Various departments teach the module utilizing a systems-based curriculum. Item difficulty, discrimination, reliability, and standard error of measurement were analyzed. Results: Item difficulty ranged from 0.3 to 0.9 on the difficulty index for most items (moderate difficulty). Most items (62/70) appropriately discriminated between high- and low-scoring students. Reliability was very high (Kuder-Richardson 20 = 0.91). The standard error of measurement was 3.7. Analysis of validity evidence included evaluation of content validity through alignment of exam items with module learning objectives using a test blueprint, as well as analysis of internal structure validity supported by item difficulty and discrimination statistics. Discrimination indices above 0.2 indicate items distinguished well between students performing at the upper and lower score ranges. Feasibility of MCQs was evidenced by the resources required. Minimal training and no specialized equipment or longer administration/scoring times were needed compared to other assessment methods. MCQs were well-accepted by students and faculty involved in test development and implementation. Conclusion: Psychometric analysis of item and exam characteristics provides validity evidence that scores from this MCQ reasonably represent CNS module achievement. While not capturing higher-order skills, MCQs proved a feasible and effective summative assessment of this pre-clinical module when used within an integrated evaluation program.

.

Corresponding author: Saeed Awad M. Alqahtani Associate professor of Physiology, Email. dr_alqahtani@hotmail.com . Tel. +966530737422

استخدام أسئلة اختيار من متعدد كوسيلة لتقييم أداء طلبة كليات الطب

مستخلص: تم استخدام المعلماتالخمسة والتي تتضمن الصلاحية والموثوقية والتكلفة والمقبولية والتأثير التعليمي - يشار إليها مجتمعة باسم "المنفعة" - لفحص "فائدة" أسئلة الاختيار من متعدد المستخدمة لقياس أداء طلاب السنة الثالثة بكلية الطب بجامعة جازان في الامتحان النهائي لمقرر القلب والأوعية الدموية وكذلك صحة الأدلة، لا سيما تلك المتعلقة بالبنية الداخلية، لدعم تفسيرات واستخدام نتائج الطلاب. استخدمت هذه الدراسة تقييما مكتوبا بالقلم والورقة بأثر رجعي لفحص أداء الطلبة باستخدام 70 سؤالا اختيار من متعدد. تم الاستشهاد بمخطط الاختبار كوثائق داعمة لموضوعية الاختبار. كما تم حساب مؤشرات الموثوقية والخطأ المعياري في القياس والصعوبة والتمييز من نتائج اختبار الطلاب. تجلت جودة الامتحان الجيدة من خلال موثوقيته العالية، وخطأ قياسي قليل في القياس، ونطاقات مناسبة لمؤشرات الصعوبة والتمييز. كان الاختبار معقولا في تكلفته ومقبولا وتعليميا. ويمكن الاستنتاج بأنه يسهل إنشاء أسئلة اختيار من متعدد عالية الجودة باستخدام "المنفعة" كإطار مرجعي مع الأخذ بالحسبان أهمية النظر في الأدلة التيتفسر نتائج الطلاب.

الكلمات الدالة:التقييم; أسئلة متعددة الخيارات; الثبات; الصدق; تحليل البنود.

Introduction

A key component of education is assessment, which tries to gauge students' aptitude for achieving set goals. Assessments might be summative or formative [1, 2]. The chance to document both strengths and places for development is provided by formative assessment, which provides feedback to students about their learning [3]. Decisions about a student's progress are generally made through summative assessment [4]. There are many kinds of evaluation techniques, some of which are subjective like the Objective Structured Clinical Exam (OSCE) and others that are objective like multiple-choice questions (MCQs) [5]. How to write effective MCQs has been detailed in numerous guidelines [6]. One of the greatest recommendations is the "Item Writing Manual," which is available on the website of The National Board of Medical Examiners [7].

Five elements, generally referred to as "Utility," including validity, reliability, acceptability, and educational impact, should be considered while developing high-quality assessments [8]. The amount to which an assessment instrument assesses what it is intended to measure is a general definition of validity [9]. However, according to the Standards for Educational and Psychological Testing, validity really refers to “the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests” [10 - 13]. Validity is typically divided into three categories: content-related, criteria, and construct validity. Nevertheless, this present perspective believes that construct validity should constitute the single, overriding paradigm used to conceive overall validity [10]. The construct validity of exam scores explanations could be backed up by evidence from a broad range of sources, including content-related, response process, internal structure, relationships to other variables, and outcomes [14, 15]. Test blueprint, often known as a table of specifications, is considered as content-related evidence to support validity [16]. Moreover, precision of measuring results is referred to as reliability which is a necessary requirement for validity [18]. There are several types of reliability including inter-rater, test-retest, parallel forms, and internal

consistency reliability which is utilized mostly with evaluation tools like MCQs [19 - 21].

Examining the test's quality after it has been administered is an essential stage in the quality assurance process, and it can be accomplished by analyzing the items using item analysis [20]. It provides a numerical evaluation of the item difficulty and item discrimination, two interconnected yet significant features of each item. The percentage of pupils who answered a question correctly is determined by the item difficulty, commonly known as the facility index [22]. Item discrimination, on the other hand, describes the proportion of right answers that differ between students with a higher overall score (the top 25%) and a lower overall score (the bottom 25%) [23]. Another crucial statistic that aids in identifying differences between the seen and true exam results is the standard error of measurement (SEM) [20]. Making decisions regarding the performance of borderline pupils is aided by SEM [16].

An evaluation method's feasibility, educational impact, and acceptability are concerned with how and when it is used. An assessment tool's feasibility is determined by weighing the resources needed to apply it against the knowledge that needs to be learned [11]. The educational impact of MCQs is seldom studied. Acceptability is the degree to which both learners and educators respond to the evaluation [11]. Instructors and learners may not appreciate the assessment if it takes too many resources to undertake, takes too long for educators to prepare, or is administered too often throughout the course [20]. MCQs are widely accepted, particularly when employed to assess the highest taxonomic levels of cognitive function [11].

The aim of the study is to evaluate the quality of the MCQs in assessing the performance of undergraduate medical students.

Methods

The preclinical phase of the system-based curriculum of the Faculty of Medicine at Jazan University

includes a central nervous system (CNS) module. This module is being taught by several departments of basic and clinical sciences. In addition to other evaluation tools, the final summative test for this module included 70 "best of five" MCQs. This study was a cross-sectional observational study with 57 students.

As was previously stated, the assessment tool's validity and reliability are a function of the scores and their interpretations rather than the assessment tool itself. The sources of evidence that can be gathered to support the construct validity of interpretations of test results for this module include content, response process, internal structure, relationships to other variables, and consequences. As proof of validity, the internal structure and validity related to the content will be discussed.

Item analysis was performed to evaluate the quality of the multiple MCQs. This involved determining how difficult each question was using difficulty index, how well it distinguished between high- and low-performing students using discrimination index. In addition, SEM was utilized to measure the error there was in the measurement of scores, and how effective the distractors were.

Due to the nature of this study, which included all students in the population, sample size calculation was not calculated.

Results and discussion

The most crucial component of written assessment, such as MCQs, is content evidence, which assesses how well a test's content corresponds to the construct it is meant to measure [10]. This proof was presented for the CNS module utilizing a test blueprint that demonstrated a clear connection between the test's learning objectives and its questions. The MCQs utilized in this examination are appropriate and reflective of the construct being looked at. The internal organization of the test is another form of validity proof. To confirm or deny evidence of the test's internal structure, analysis of the test's items to determine item discrimination and item difficulty indices and calculation of the reliability are required6.

Item analysis of the test

There are 70 questions, and each one receives 1 mark for the right response and 0 for the wrong one. The data is regularly distributed, as evidenced by the mean and median scores, which are nearly equal at 42.79 and 42, respectively (appendix A). Because mean can be used to represent a real central tendency, standard deviation (SD) represents the distribution of the data (students' scores).

The percentage of pupils that correctly respond to an item is referred to as the item's difficulty. The appropriate value for the difficulty index should fall between 0.3 and 0.9, per the assessment requirements for the faculty of medicine at Jazan University. Only 7% of the exam items are challenging since their difficulty (facility) index is less than 0.3, while the remaining 4% are simple because their difficulty index is greater than 0.9. The difficulty index for the remaining MCQs ranges from 0.3 to 0.9.

The discrimination index is calculated using upper and lower 25% of each item's values (upper and lower 14 scores). Most of the MCQs (62 out of 70 items) have scores that discriminate between students who performed well and those who did not perform well in an acceptable manner; their item discriminations are greater than 20%. A discrimination index of less than 20% is seen for eight items. Five of these questions are not discriminating because they are either very difficult or very easy. Despite measuring vital content, they are the least successful psychometrically, which increases the test results' validity in relation to the content they cover. The remaining three (items 5, 28, and 43) have a negative discrimination index and fall within the 0.3–0.9 acceptable difficulty range. Before including them in the questions bank, these items need to be fixed and may even be deleted or drastically altered. The incorrect interpretation of the question by high achievers may be to blame for the negative discrimination, or the item may have given a hint to low achievers who were able to figure out the right response. There are 0.61 and 0.46 mean difficulty and discrimination indices, respectively. This indicates that the test is moderately challenging and highly discriminative of high- and low-performing pupils. The statement that "the most

informative test items are those of moderate difficulty which discriminate highly" has led to the classification of this exam as fair and acceptable. Appendix B contains a selection of the MCQs from this test.

As evidenced by the reliability coefficient (KR-20) of 0.91, test results are very reliable. It is even higher than the suggested range (0.8–0.89) for tests with moderate stakes, such end-of-year summative exams in medical school [10]. The SEM aids in decision-making regarding the performance of borderline students and in constructing confidence intervals around observed test scores [12,13]. The minimum passing score for this test is 60, and the SEM is 3.7. This translates to a 68% confidence level among the examiners that the student's actual scores should fall between 54.3 and 63.7. For students who score between 60 and 63.7, other activities should be considered when determining whether the student passed the exam. Appendix A provides more details and illustrations regarding this test item analysis.

In general, this test evaluates the Miller's pyramids "knows" and "knows how" levels, Figure 1. The test's items are also tied to various cognitive domain levels according to Bloom's cognitive taxonomy for educational goals.

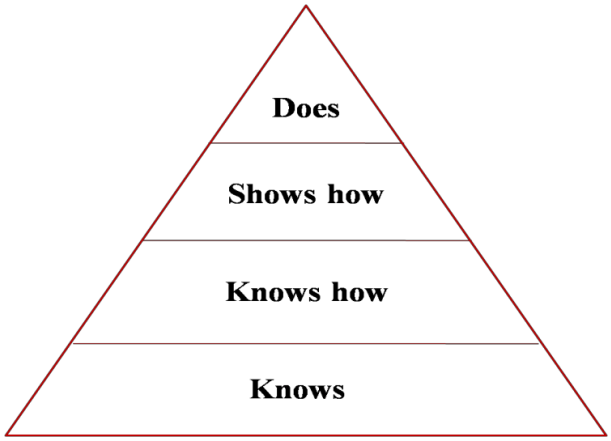

Figure 1: Miller’s Pyramid of Competence. Based on work by Miller 23.

Feasibility, Accessibility and Educational Impact

There was no specific training required for this test. Content specialists from the several departments participating in teaching the CNS module created each MCQ that was utilized. To ensure that they are written in accordance with the standards, these items are updated by a committee called the student assessment committee. Additionally, after receiving the results from the students, this committee updated the exam. It did take the professors a long time to prepare these products. The findings were provided in a matter of minutes after the students used response sheets to complete the multiple-choice questions. In fact, MCQs are less expensive and require less resources and time for adjustments than other methods like short answer questions and modified essay questions. MCQs were well-received and accepted as an assessment method by both the students and the faculty members who helped with test preparation.

Measuring the effect of education is incredibly uncommon. However, it is generally

acknowledged that learning is driven by assessment. As a result, the examiner assumed that pupils had learned something by studying for and taking the test. Additionally, students receive feedback on their performances, and the input they provided on the test and the entire module had a beneficial educational influence.

The findings of this study may be applicable to other medical schools and other educational institutions that use MCQs to assess student learning. However, it has several limitations. First, the sample size was relatively small (N = 57) and was conducted at a single institution, which limits the generalizability of the findings.

Conclusion

Various medical departments teach the CNS module. Along with other techniques for evaluation, MCQs are used to gauge students' performance. The Student Assessment Committee updated the exam to ensure that the items were written in accordance with the rules. It was noted where the sources of support for the construct validity came from. Utilizing the test blueprint and computing the reliability coefficient (KR-20) of the internal consistency, which is 0.91, respectively, allowed for the examination of evidence such as content-related validity and internal structure. The validity of test results is supported by this evidence. The test's internal structure is also supported by the item difficulty and item discrimination factors. These test items often have a medium level of difficulty and are quite discriminating. In a similar vein, these characteristics of item difficulty and item discrimination are what Downing recommended for informative test items in his book "evaluation of health professions education" [17]. The SEM is a helpful tool for applying dependability coefficient to show the precision of measurement and the degree of measurement error [13]. For the student who scored between 60 and 63.7 on this test, subsequent activities should be evaluated based on the SEM. The test is generally reliable, highly acceptable, and has a positive impact on education.

Appendix A

Item Analysis – Central Nervous System Module

Faculty of Medicine, Jazan University

Medical Education Unit

Students Assessment

Committee (SAC) MCQs Item

Analysis Report Form

General Analysis:

- Highest score: 63
- Lowest score: 20
- Mean: 42.79
- Median: 42
- Standard Deviation (SD): 11.94
- Reliability Coefficient (KR20): 0.91
- Standard Error of Measurement: 3.7
- Mean difficulty index: 0.61
- Mean discrimination index: 046
- Total number of questions: 70
- Cut off score: 42 (60%)
- Total number of examinees: 57

		Item number	Total No.		Percent
Item difficult	Items with Difficulty Index <0.3	10,30,42,46,67	5		7%
	Items with Facility Index >0.9	7,40,57	3		4%
tem discrimination	Items with –ve discrimination	5,10,28,30,42,43,67	7	8	12%
	Items with zero discrimination	-	-		
	Items with Discrimination Index <0.2	40	1		
	Items with Discrimination Index≥0.2	1,2,3,4,6,7,8,9,11,12,13,14,15,16,17,18,19,20,21,23,24,25,26,27,29,3	62		88%
Distracters	Items with chosen distracters more than	10,30,42,46,67	5		7%
	Items with one option not used	2,3,6,13,14,22,27,62,69	9	14	20%
	Items with two options not used	7,50,56	3		
	Items with more than two options not	40,66	2		

Comments:

1) Items with –ve discrimination should be detected **or** radically changed.

2) Items with zero discrimination should be revised for its importance and decide accordingly.

3) Items with problems in the options should be revised and discussed with SAC before included in the Bank.

4) Questions with multi-defects should be deleted.

5) Questions where the students have chosen options more than the key answers should be revised in the department and then with SAC.

Appendix B

Sample of the MCQs – Central Nervous System Module

27. Ahmed is a 60-year-old suffers from muscle weakness with loss of deep and fine tactile sensations. Few weeks later, crude sensations recover and accompanied by emotional disturbances. Which of the following diagnosis is appropriate?

A) Tabes dorsalis. B) Syringomyelia.

C) Subacute combined degeneration.

D) Thalamic syndrome. E) Neocerebellar lesion.

Correct answer: D Item difficulty: 0.9

Item discrimination: 0.3

30. A 45-year- old worker exposed to sever head injury. CT of brain revealed cranial hematoma. By clinical examination, there was normal fine precise movements of both hands but he suffered loss of 2 point discrimination, where was the provisional site of hematoma?

A) Premotor cortex.

B) Anterior limb of internal capsule. C) Corona radiata.

D) Pontine tegmentum. E) Medullary pyramid.

Correct answer: B Item difficulty: 0.18

Item discrimination: -0.2

40. Which of the following conditions is caused by irregular curvature of the cornea?

A) Presbyopia. B) Astigmatism.

C) Hypermetropia. D) Emmetropia.

E) Myopia.

Correct answer: B Item difficulty: 0.93
Item discrimination: 0.1

46. **A 50-year old male of low socioeconomic status was presented to the hospital complaining of headache and blurred vision, his clinical examination revealed fever and neck retraction. Subsequent CSF examination revealed turbidity, reduced glucose, increased protein, and lymphocytes 400 cells/mm^3, which of the following diagnostic methods is used for identification of the causative microorganism responsible for this condition?**

A) Oxidase test. B) Gram staining.
C) Ziehl-Neelsen staining.
D) Culture on chocolate gar.
E) Test of acid production from sugars fermentation.
Correct answer: C Item difficulty: 0.21
Item discrimination: 0.4

57. Which of the following is a feature regarding the morphology of brain abscess?

A) Excess infiltration by neutrophils. B) Normal convolutions.
C) No brain tissue destruction. D) Healing by regeneration.
E) No cyst formation.
Correct answer: A Item difficulty: 0.91
Item discrimination: 0.2

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 author has no conflict of interest to declare.

Ethical Approval
There was no ethical issue requiring approval from an institutional review board for this study, as it involved secondary analysis of anonymous examination data routinely collected for quality assurance purposes, with no involvement of human subjects.

Author contributions
The author declares that he is the sole contributor to this work

References:

1. **Bhattacharya S, Mondal S, Mitra K.** Performance of Undergraduate Medical Students in Formative and Summative Evaluations in Community Medicine in a Medical College in India. Journal of Medical Education. 2021;20(4).
2. **Al-Wardy NM**. Assessment methods in undergraduate medical education. Sultan Qaboos University Medical Journal. 2010;10(2):203.
3. **Nicol D, Macfarlane-Dick D**. Rethinking formative assessment in HE: a theoretical model and seven principles of good feedback practice. C Juwah, D Macfarlane-Dick, B Matthew, D Nicol, D & Smith, B(2004) Enhancing student learning though effective formative feedback, York, The Higher Education Academy. 2004;
4. **Cilliers FJ, Schuwirth LW, Herman N, Adendorff HJ, van der Vleuten CP**. A model of the pre-assessment learning effects of summative assessment in medical education. Advances in Health Sciences Education. 2012;17(1):39–53.

5. **Vincent SC, Arulappan J, Amirtharaj A, Matua GA, Al Hashmi I.** Objective structured clinical examination vs traditional clinical examination to evaluate students' clinical competence: A systematic review of nursing faculty and students' perceptions and experiences. Nurse Education Today. 2022;108:105170.
6. **Stevens SP, Palocsay SW, Novoa LJ**. Practical Guidance for Writing Multiple-Choice Test Questions in Introductory Analytics Courses. INFORMS Transactions on Education. 2022;
7. **Adnan S, Sarfaraz S, Nisar MK, Jouhar R**. Faculty perceptions on one-best MCQ development. The Clinical Teacher. 2022;e13529.
8. **Naeem N.** Validity, reliability, feasibility, acceptability and educational impact of direct observation of procedural skills (DOPS). J Coll Physicians Surg Pak. 2013;23(1):77–82.
9. **Sullivan GM**. A Primer on the Validity of Assessment Instruments. J Grad Med Educ. 2011 Jun;3(2):119–20.
10. **American Educational Research Association**. Standards for educational and psychological testing. American Educational Research Association; 2018.
11. **Yudkowsky R, Park YS, Downing SM**. Introduction to Assessment in the Health Professions. In: Assessment in health professions education. Routledge; 2019. p. 3–16.
12. **Lane S, Raymond MR, Haladyna TM, Downing SM**. Test development process. In: Handbook of test development. Routledge; 2015. p. 19–34.
13. **Downing SM, Juul D, Park YS**. Statistics of testing. In: Assessment in health professions education. Routledge; 2019. p. 70–85.
14. **Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R**. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Advances in Health Sciences Education. 2014;19(2):233–50.
15. **Arjoon JA, Xu X, Lewis JE**. Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence. Journal of Chemical Education. 2013;90(5):536–45.
16. **AlFallay IS**. Test specifications and blueprints: Reality and expectations. International journal of instruction. 2017;11(1):195–210.
17. **Downing SM**. Validity: on the meaningful interpretation of assessment data. Medical Education. 2003;37(9):830–7.
18. **Palmen LN, Kosse NM, van Hooff ML, Witteveen AG.** Evaluation and Validation of the Dutch European Foot and Ankle Society (EFAS) Score. The Journal of Foot and Ankle Surgery. 2022;61(3):464–70.
19. **Cook DA, Beckman TJ**. Current concepts in validity and reliability for psychometric instruments: theory and application. The American journal of medicine. 2006;119(2):166-e7.
20. **Mavis B.** Assessing Student Performance. In: Jeffries WB, Huggett KN, editors. An Introduction to Medical Teaching [Internet].

Dordrecht: Springer Netherlands; 2010 [cited 2022 Nov 10]. p. 143–78. Available from: https://doi.org/10.1007/978-90-481-3641-4_11

21. **Brenner E.** Anatomy in Competencies-Based Medical Education. Education Sciences. 2022;12(9):610.
22. **Baldwin BA**. The role of difficulty and discrimination in constructing multiple-choice examinations: With guidelines for practical application. Journal of Accounting Education. 1984 Mar 1;2(1):19–28.
23. **Miller MD**, Linn RL, Gronlund NE. Measurement and Assessment in Teaching. Merrill/Pearson; 2009. 551 p.