ORIGINAL ARTICLE

MULTIPLE CHOICE QUESTIONS IN ASSESSING MEDICAL STUDENTS' ACHIEVEMENTS IN ENDOCRINOLOGY MODULE IN REFERENCE TO 'UTILITY'

Saeed A. Al-Qahtani

Department of Physiology, Faculty of Medicine, Jazan University, Jazan City, Saudi Arabia

Background: Usefulness of the MCQs was defined in relations to five factors: validity, reliability, cost, acceptability and educational impact; collectively called 'utility'. The aim of this study was to use 'utility' as a frame of reference to assess the usefulness of the multiple choice questions to measure the performance of 3rd year medical students, Jazan University, in the final exam of Endocrinology module (2013/2014), also to evaluate the validity evidences especially those related to internal structure to support the interpretations and the use of the students' results. **Methods:** This study was a retrospective of written pen and paper assessment tool utilizing a sample of 62 of 3rd year medical students in the final exam of Endocrinology module. Seventy single best answer MCQs were used. Test blueprint used as a source of content-related evidence. Students' results were analysed to measure reliability and standard error of measurement, difficulty and discrimination indices. **Results:** The quality of the exam was high as indicated with the high reliability, low standard error of measurement and with values of difficulty and discrimination indices in acceptable ranges. The exam was not costly, acceptable, and had an educational impact. **Conclusion:** Using utility as a reference frame helps in producing high quality MCQs. However, evidences that support the interpretations and use of the students' results should be considered.

Keywords: Multiple choice questions, utility, validity evidences

Pak J Physiol 2016;12(3):23–6

INTRODUCTION

Five factors should be considered to evaluate the quality of the MCQs which are validity, reliability, cost, acceptability and educational impact; collectively called 'Utility' as first proposed by Van Der Vleuten in 1996.[1,2] Validity refers to 'the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests' as stated in the 6th version of Standards for Educational and Psychological Testing (Standards).[3]

Compared to the classifications of validity stated in previous four versions of standards, the 5th and 6th versions saw that all types of validity should be represented only by construct validity.[3] Subsequently, examples of evidences from different sources should be provided to support the validity.[4] These sources include test content[5], response process[6], relations to other variables[7], consequences[8] and internal structure[9]. Reliability is referred to the consistency or repeatability of measurement scores.[3] As they have an objective scoring process, MCQ test's scores have a high degree of reliability.[10] However, there are many factors that affect this reliability such as timing, length and construction of the test and some environmental factors.[11,12]

There are four general types of reliability which include internal consistency, parallel forms, test-retest and inter-rater reliabilities.[13,14] Internal consistency is estimated by assessing the correlation across items for a single test administered once.[15,16] Therefore, it is logically used most commonly with an assessment tool such as MCQs. There are many internal consistency subtypes including Kuder-Richardson Formula 20 (KR-20) which is used if the test items are scored dichotomously, thus, it is appropriate reliability estimate for MCQ test's scores.[15,17]

As part of the quality assurance, examining the quality of the test after its administration is a crucial step which can be achieved by analyzing the items' characteristics.[17,18] It provides a numerical assessment of two interrelated important characteristics of each item, which are item difficulty and item discrimination. Both are considered as source of information for internal structure of the exam —an evidence for the validity as mentioned above. Item difficulty, also called facility index, determines the percentage of students who answered an item correctly.[19] Item discrimination refers to the proportion difference in correct responses between students with a higher overall score (the top 25%) and lower overall score (the bottom 25%).[19] The standard error of measurement (SEM) is another important statistic which helps to determine the discrepancies between the observed and true scores of the examinees.[20] It helps in decision making about the performance of borderline students.[12]

Cost is a decision making about the required recourses to implement the assessment tool.[2] Educational impact is a source of information related to consequences evidence.[21,22] Acceptability is the extent

of responsiveness of the students and teacher to the assessment.[2]

Endocrinology module (END) is taught in the preclinical phase of the system-based curriculum in the Faculty of Medicine, Jazan University, KSA. Several basic and clinical sciences departments are involved in teaching this module. Best of five MCQs were used among other assessment instruments in the final summative test for this module. The aim of this study was to assess the students' achievement in terms of the 'utility' and to evaluate the validity evidences especially those related to internal structure to support the interpretations and the use of the students' results.

METHODS

This was a retrospective study of written pen and paper assessment tool utilizing a sample of 62 3rd year female medical students of Jazan University, Jazan City, KSA during the 2013/2014 academic year taking final summative exam for END module with the permission of the Ethical Committee. Seventy single best answer MCQs were used with five options, four alternatives and one answer. The MCQs were constructed in accordance with the item-writing guidelines of the faculty of Medicine. The MCQ items were first written or extracted from the bank by individual teachers involved in the teaching of this module, then collected by the module coordinator to be submitted to the Student Assessment Committee. This committee, includes most experienced academic staff in the faculty, and works on content accuracy and relevance to the module objectives using the blueprint —a source of information for content-related evidence. The students answered the MCQs using answer sheet and they were checked using special machine. In scoring the tests, each item got '1' mark for the correct answer and '0' for incorrect answer. Analysis of the students' scores was conducted to assess the difficulty and discrimination indices, reliability and standard error of measurement.

The average of the students' results was expressed as Mean±SD. The reliability and the standard error of measurement of the exam and the difficulty and discrimination indices of each item were calculated using SPSS.

RESULTS

The cut-off score for failing in this test was 42 (60%). The scores mean was 47 and the standard deviation was 10. In this test only 4.3% of items were very difficult as their difficulty (facility) index was less than 0.3 while 13.6% were very easy; with difficulty index more than 0.9. Further, fifty-eight items (around 82%) had moderate difficulty index between 0.3 and 0.9. However, there of these 58 items had low and 2 had negative discrimination indices.

The discrimination indices of most of the MCQs (60 out of 70 items) were more than 0.2. However, six of the 60 items had high and one had low difficulty indices. Only seven items had low discrimination index (between 0.0 and 0.2). Three of them were within the recommended range of difficulty. However, the other four were either very difficult (one item) or very easy (3 items). Further, three items had negative discrimination index; one item was very difficult and two were in the acceptable range of difficulty.

A total of 53 items were moderately difficult and had high discrimination index, Table-1. The means of the difficulty and discrimination indices were 0.7 and 0.4, respectively.

Table-1: The items distribution according to the levels of difficulty and discrimination indices

	Difficulty		
Discrimination	High (9 items)	Moderate (58 items)	Low (3 items)
High (60 items)	6	53	1
Low (7 items)	3	3	1
Negative (3 items)	0	2	1

The reliability coefficient, which was represented in the test by internal consistency estimate, and measured by KR-20, was 0.89 and the SEM was 3.3. Detailed information about the items analysis is illustrated in Table-2.

Unlike other tools like short answer questions and modified essay questions, indeed, MCQs were not costly and did not need a lot of recourses and time for corrections. Both students and the faculty involved in teaching this test were satisfied and accepted MCQs as an assessment tool.

The examiners assumed that students learnt by preparing and undertaking this test. In addition, feedback was given to students about their performances and the student gave feedback about the test which is a usual process after each exam.

DISCUSSION

Validity and reliability are related primarily to students' scores, their interpretations and uses.[3] Both are important factors that affect the exam quality. Mainly, the internal structure is discussed as validity evidence for a module. The scores mean was above the cut-off score of failing in this test. Most items, as recommended by the guidelines of assessment in the Faculty of Medicine in Jazan University and other guidelines, were moderately difficult ranging from 0.3 to 0.9.

Top and bottom 25% of scores (top and bottom 16 scores) of each item were used to compute the discrimination index. The scores of the most items (61 out of 70 items) were acceptably discriminating between students with good performance and students

who did not perform well; their item discriminations were more than 0.2. However, 10% of the items had low discrimination index (between 0.0 and <0.2), where around half of them (4 out of these 7 items) contributed to this low values as they are not within the recommended range of difficulty. Therefore, these 4 items were the least effective psychometrically, but measuring important content, thus, enhancing the content-related validity of the test scores. Furthermore, low performing students answered 3 items whereas high performing students did not as they had negative discrimination index. These items were flawed and should be revised and may be deleted (ones with low difficulty index) or radically changed (ones with moderate difficulty) before adding them to the questions bank. The negative discrimination may, due to misinterpretation by high performing students or the item, provided a clue to low performing students enabling them to guess the correct answer. Fifty-three items can be added to questions bank with no further modification as they had high discrimination index with moderate difficulty. The mean difficulty and discrimination indices were 0.7 and 0.4, respectively. This means that the test had moderate difficulty and high discrimination between high and low performing students. This test was considered fair and acceptable as it was stated that 'the most informative test items were those of middle difficulty which discriminate highly'.[13]

The reliability was high, as the reliability coefficient (KR-20) was 0.89. It was within what is recommended for moderate stakes tests (0.8–0.89), such as end-of-year summative tests in medical school.[16] The SEM helps in building confidence intervals around observed test scores and in making decision about the performance of borderline students.[12] The cut-off score for failing in this test was 42 (60%) and the SEM was 3.3. This means that the examiners were 68% confident that the students' true scores should lie between her observed score ±3.3. Therefore, other activities should be considered for students who score between 42 and less than 45.3 to decide whether the student pass the test.

This test did not need special preparation. All used MCQs were constructed by content experts from the different departments involved in teaching the END module. These items were revised by a committee called Student Assessment Committee to make sure that they were written according to the guidelines. In addition, this committee revised the test after the students results was obtained. Indeed preparing these items took long time from faculty. The students answered the MCQs using answer sheet and they were checked using special machine and the results were ready in minutes. Unlike other tools like short answer questions and modified essay questions, indeed, MCQs were not costly and did not need a lot of recourses and time for corrections. Both students and the faculty involved in teaching this test were satisfied and accepted MCQs as a good assessment tool.

It is extremely rare that educational impact is measured. However, it is widely accepted that assessment drives learning. Based on that the examiners assumed that students learnt by preparing and taking this test. In addition, feedback was given to students about their performances and the student gave feedback about the test which is a usual process after each examination. Further, using a cut-off score for failing in this test gave useful information to support the consequence evidence.

CONCLUSION

Using utility as a reference frame helps in producing high quality MCQs. However, evidences that support the interpretations and use of the students' results should be considered.

REFERENCES

1. Al-Wardy NM. Assessment methods in undergraduate medical education. Sultan Qaboos Univ Med J 2010;10(2):203–9.
2. Van Der Vleuten CP. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theory Pract 1996;1(1):41–67.
3. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US). The Standards for educational and psychological testing. USA: American Educational Research Association; 2014.
4. Rios J, Wells C. Validity evidence based on internal structure. Psicothema 2014;26(1):108–16.
5. Sireci S, Faulkner-Bond M. Validity evidence based on test content. Psicothema 2014;26(1):100–7.
6. Padilla JL, Benítez I. Validity evidence based on response processes. Psicothema 2014;26(1):136–44.
7. McCoach DB, Gable RK, Madura JP. Evidence Based on Relations to Other Variables: Bolstering the Empirical Validity Arguments for Constructs. In: DB McCoach, RK Ga, JP M, (Editors). Instrument Development in the Affective Domain. USA: Springer; 2013. pp. 209–48.
8. Lane S. Validity evidence based on testing consequences. Psicothema 2014;26(1):127–35.
9. Downing SM, Haladyna TM. Validity and its threats. In: Downing SM, Yudkowsky R, (Editors). Assessment in health professions education. New York: Routledge; 2009. pp. 21–56.
10. Gupta M, Sharma G, Pal RAG, Thaman R, Tikoo D. Strategic use of MCQs in undergraduate medical students to improve objectivity of formative assessment. Natl J Integr Res Med 2012;3(2):113–8.
11. Cohen RJ, Swerdlik ME, Sturman E. Psychological testing and assessment: An introduction to tests and measurement. 8th ed. New York: McGraw-Hill; 2012.
12. Tavakol M, Dennick R. Post-examination analysis of objective tests. Med Teach 2011;33:447–58.
13. Downing SM. Statistics of testing. In: Downing SM, Yudkowsky R, (Editors). Assessment in health professions education. New York: Routledge; 2009. pp. 93–118.
14. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006;119(2):166.e7–e16.
15. DeVellis RF. Scale Development: Theory and Applications. USA: SAGE Publications; 2012.
16. Axelson R, Kreiter C. Reliability. In: Downing SM, Yudkowsky R, (Editors). Assessment in health professions education. New York: Routledge; 2009. pp. 57–74.

17. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004;38:1006–12.
18. Siri A, Freddano M. The use of item analysis for the improvement of objective examinations. Procedia Soc Behav Sci 2011;29:188–97.
19. Collins J. Education techniques for lifelong learning: Writing multiple-choice questions for continuing medical education activities and self-assessment modules. Radiographics 2006;26(2):543–51.
20. Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: is there a connection? J Biopharm Stat 2004;14(1):97–110.
21. Brian M. Assessing student performance. In: Jeffries WB, Huggett K, (Editors). An introduction to medical teaching. USA: Springer; 2010.
22. Van Der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005;39:309–17.

Address for Correspondence:

Dr. Saeed A. Al-Qahtani, Department of Physiology, Faculty of Medicine, Jazan University, Jazan City, Saudi Arabia.
Email: dr_alqahtani@hotmail.com

Received: 25 Jun 2016 Revised: 16 Sep 2016 Accepted: 22 Sep 2016

Appendix-1: Endocrinology Module –General Item Analysis

Scores Mean: 47.1 **SD**: 10 **Scores Median**: 48.5			**Cut-off score**: 42 (60%) **Reliability Index**: 0.89 **SEM**: 3.3			**Total number of Items**: 70 **Dif Mean**±SD: 0.7±0.2 **Dis Mean**±SD: 0.4±0.2		
Item #	**Dif**	**Dis**	**Item #**	**Dif**	**Dis**	**Item #**	**Dif**	**Dis**
1	0.95	0.2	24	0.34	0.2	47	0.92	0.1
2	0.77	0.3	25	0.83	0.4	48	0.95	0.2
3	0.81	0.3	26	0.72	0.6	49	0.31	0.2
4	0.53	0.8	27	0.45	0.5	50	0.92	0.3
5	0.78	0.4	28	0.89	0.3	51	0.91	0.4
6	0.91	0.3	29	0.73	0.5	52	0.48	0.6
7	0.50	0.5	30	0.86	0.4	53	0.64	0.4
8	0.67	0.5	31	0.64	0.8	54	0.53	0.5
9	0.41	0.4	32	0.94	0.3	55	0.72	0.2
10	0.88	0.4	33	0.55	0.2	56	0.89	0.2
11	0.80	0.6	34	0.73	0.6	57	0.42	0.4
12	0.73	0.4	35	0.56	0.7	58	0.58	0.1
13	0.64	0.3	36	0.73	0.6	59	0.66	0.5
14	0.83	0.5	37	0.78	0.4	60	0.30	-0.1
15	0.48	0.6	38	0.42	0.2	61	0.41	0.1
16	0.75	0.7	39	0.33	0.4	62	0.91	0.3
17	0.83	0.5	40	0.64	0.1	63	0.95	0.1
18	0.83	0.5	41	0.33	0.4	64	0.73	0.5
19	0.88	0.4	42	0.97	0.1	65	0.19	-0.1
20	0.75	0.4	43	0.45	0.4	66	0.80	0.4
21	0.23	0.1	44	0.63	0.5	67	0.59	0.2
22	0.47	0.3	45	0.70	0.6	68	0.69	0.6
23	0.53	-0.3	46	0.86	0.3	69	0.81	0.4
						70	0.72	0.4

SEM=Standard error of measurement, Dif=Difficulty, Dis=Discrimination