Open Access

# Psychometric Characteristics of MCQs used in Assessing Phase-II Undergraduate Medical Students of Universiti Sains Malaysia

A. Barman*[,1], R. Ja'afar[2], F.A. Rahim[3] and A.R. Noor[4]

[1]*Department of Medical Education, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kota Bharu, Malaysia*

[2]*Department of Medical Education, School of Medical Sciences, Universiti Sains Malaysia*

[3]*Department of Medical Education, School of Medical Sciences, Universiti Sains Malaysia*

[4]*Academic and Student Development, School of Medical Sciences, Universiti Sains Malaysia*

**Abstract:** Multiple choice questions (MCQs) are used in assessing both undergraduate and postgraduate medical students of the School of Medical Sciences (SMS), Universiti Sains Malaysia (USM). Questions that are generated by the subject specialists are vetted at departmental and central levels. MCQs used in assessing Phase-II MD students are analysed in terms of its reliability, validity and difficulty and discriminating indices. For reliability in terms of internal consistency both Spearman-Brown formula and Cronbach's alpha were used. Difficulty and discriminating indices for the MCQs were collected from the computer generated marking sheets. Alpha reliability coefficient for internal consistency is 0.91 for both MCQ1 and MCQ2 while corrected reliability (Spearman-Brown prophecy) for MCQ1 is 0.88 and MCQ2 is 0.91. The process of generating and vetting of questions judges the face and content validity of both these MCQs. The concurrent validity for MCQ1 and MCQ2 are r=0.55, p<0.01 and r=0.69, p<0.01 respectively. Sixty percent of both MCQ1 and MCQ2 are within the difficulty index of 20% to 80%, while 34% of MCQ1 and 37% of MCQ2 have discriminating indices of ≥0.2. The MCQs have satisfactory levels of reliability and validity. A majority of the MCQs are within the acceptable level of difficulty index. A well-structured and strict central vetting process in the SMS ensures an acceptable standard of MCQs.

**Keywords:** Psychometric characteristics, MCQs, undergraduate medical students.

## INTRODUCTION

For every evaluative action, there is an equal if not greater and sometimes opposite educational reaction [1]. Proper selection of assessment methods can improve student performance [2]. Along with quality assurance in all its activities, the School of Medical Sciences, Universiti Sains Malaysia is serious about the quality assessment of its students.

MCQs are used in assessing both undergraduate and postgraduate medical students of the School of Medical Sciences (SMS), Universiti Sains Malaysia (USM) along with the essay, MEQs, OSCEs, short case and long case. The MCQ is one of the most established objective, reliable and valid methods of assessment, but all these depend on how carefully the test items are prepared. In SMS, questions that are generated by the subject specialists are vetted at departmental and central levels in terms of their relevancy to learning objectives, wording and structuring. This study is intended to analyse MCQs used in assessing Phase-II MD students of USM in terms of its reliability, validity and difficulty and discriminating indices. Though SMS uses both multiple true-false and single best response type of MCQs, here only the multiple true-false type MCQs were analysed.

## METHODS

A total of 180 students sat for the second professional examination of April 2004 with a pass rate of 93.3%. They attempted 100 MCQs each with 5 true-false responses. Marks obtained by the individual student on each of the MCQs of MCQ1 and MCQ2 and MEQ1 and MEQ2 as a whole were collected and entered in the SPSS computer programme.

For reliability in terms of internal consistency, both Spearman-Brown formula and Cronbach's alpha were used. The concurrent validity was assessed by Pearson correlation between the MCQs and respective MEQs (Annex 1). No specific test was done for face and content validity. Difficulty and discriminating indices of the MCQs were collected from the computer generated marking sheet available in the academic office. Difficulty index is one of the most frequently reported, item analysis statistics. It is a measure of the percentage of examinees who answered the item correctly. Equation used to calculate it is:

$$\mathrm{DI} = \frac{\text{Number of students make the answer correct for the item}}{\text{Number of students attempted the item}} \times 100$$

This is a measure of how difficult the item is to answer by the students correctly. The higher is the value of DI the easier the item is.

Discriminating index or power is a measure of how well an item is able to distinguish between examinees who are

*Address correspondence to this author at the Department of Medical Education, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kota Bharu, Malaysia; Tel: +6 09 7664112; Fax: +6 09 7653370; E-mail: barman@kb.usm.my

1876-519X/10 2010 Bentham Open

knowledgeable and those who are not. Item discriminating power is calculated by subtracting the number of examinee in the lower group (low total score) who make the item right from the examinee of upper group (high total score) who make the item right and dividing by the number of students in one group. Use of upper and lower subgroups each containing 27 per cent of the total examinee is quite common in item analysis [3-5]. If there are 100 examinees and 15 examinees of upper group(27) make the item correct and 5 examinees of lower group(27) make the item correct then, discriminating power or index is: D = (15-5)/27 = 0.37.

$$\text{Dis. Index} = \frac{\begin{array}{c}\text{No. of examinee in the upper}\\ \text{group who make the item correct}\end{array} - \begin{array}{c}\text{No. of examinee in the lower}\\ \text{group who make the item correct}\end{array}}{\text{No. of examinee in one group}}$$

An item with discriminating power 1 indicates that all students in upper group make the item right and all students in lower group make the item wrong. An item with Zero discriminating power means equal number of students in upper and lower groups make the item right.

## RESULTS

Reliability in terms of internal consistency is analysed for MCQ1 and MCQ2 using split-half method as well as alpha reliability coefficient. Spearman-Brown prophecy formula was used in split-half method for corrected reliability. Alpha reliability coefficient is 0.91 for both MCQ1 and MCQ2 while corrected reliability (Spearman-Brown prophecy) for MCQ1 is 0.88 and MCQ2 is 0.91.

No specific statistical test was done for face and content validity. The face validity and content validity of both these MCQs are reasonably ensured by the systematic process of generating and vetting of questions.

Sixty percent of both MCQ1 and MCQ2 are within the difficulty index of 20% to 80% (Table **1**), while 34% of MCQ1 and 37% of MCQ2 have discriminating indices of 0.20 and above (Table **2**).

## DISCUSSION

The MCQs used in the professional II examination of the MD programme have satisfactory levels of reliability and validity. Reliability coefficient for internal consistency by using Spearman-Brown formula is 0.88 for MCQ1 and 0.91 for MCQ2. Buckley-Sharp *et al.* 1972 [6] stated that the value varies from 0.06 to 0.95 based on the number of test items and use of the results. Usually the ranking test with larger number of test items need higher reliability coefficient. Cronbach's alpha of 0.91 for both MCQ1 and MCQ2 suggest that the test items were reliable. Tests commonly have reliability coefficient between 0.60 and 0.85 [7-11].

Face and content validity measure how well the test items represent the domain of learning objectives. There is no statistics to establish the content validity [12]. Test is judged to have content validity as it is designed and evaluated by expert faculty [13, 14]. The consensus development techniques justify both face and content validity [15]. A well-structured process of generating and vetting of questions in SMS, USM ensured the face and content validity. Pearson's correlation coefficient of .547 (MCQ1 and MEQ1) and .691 (MCQ2 and MEQ2) are reasonably acceptable to support the concurrent validity (Annex1). Items were not verified to see if there is any difference of measurement domain by MCQs and MEQs that may have some influence on the low levels of correlation. Majority of the MCQs are within the acceptable level of difficulty index, which measures the percentage of students who got the item right. Dixon, 1994 [16] advocates the diffi-

**Table 1. MCQ1 and MCQ2 by Difficulty Index**

| Difficulty Index | No. of MCQ1 Items (%) | No. of MCQ2 Items (%) |
|---|---|---|
| <0.20 | 18 (3.6%) | 21(4.2%) |
| .20 to .80 | 302 (60.4%) | 298(59.6%) |
| >0.80 | 170 (34.0%0 | 181(36.2%) |
| Total | 500 (100)% | 500(100.0%) |

**Table 2. MCQ1 and MCQ2 by Discriminating Index**

| Discriminating Index | No. of MCQ1 Items (%) | No. of MCQ2 Items (%) |
|---|---|---|
| <0.20 | 330 (66.0%) | 312(62.8%) |
| =>0.20 | 170 (34.0%) | 188 (37.2%) |
| Total | 500 (100.0%) | 500 (100.0%) |

culty index of 20-80% for multiple true-false MCQs. The difficulty index is not solely determined by the content of the item as it also reflects the ability of the examinee [17] and the instruction they have had [7]. For a well-prepared group of examinees item difficulty indices may range from 70 to 100% [17]. In the second professional examination a 93.3% pass rate indicates a well-prepared group of students and this may be the reason of high difficulty indices for some of the test items. A rigid content specification should be maintained in generating the items [17] and for that purpose, items with high difficulty indices may need to be accepted. MCQ items with good discriminating potential tend to be moderately difficult items, and the moderate to very difficult items are more likely to have negative discrimination [18].

Discriminating index refers to the degree to which the test item discriminates between students with high and low achievement. When the difficulty index moves towards high or low from 50%, the discriminating index becomes low [7]. Preferable discriminating indices are 0.20 and above [16], but in criterion-referenced measurement, many good items may have discrimination indices of zero [17].

True-false format MCQs provide cues, resulting in a less discriminatory index [19]. Again low discriminating index is more likely if the test measures a variety of types of learning outcomes. For validity, a well-constructed test accepts items with low discriminating indices [7]. The so-called 'assessment by ambush' is one aspect of unfair examination, where, for high discrimination, potentially important areas are not tested [20]. With regards to the School of Medical Sciences curriculum, questions are integrated and measure a variety of learning outcome based on criterion-reference. These may be the reasons for the low discriminating index for many of the test items.

## CONCLUSION

The MCQs used in the professional II examination of the MD programme have satisfactory levels of reliability and validity. MCQs are within the acceptable level of difficulty index. A well-structured and strict central vetting process in the medical school helps to ensure an acceptable standard of MCQs.

## ACKNOWLEDGEMENTS

The authors would like to thank the Academic Department and the Computer Unit of SMS for their help in information collection and the computer software.

**Annex 1: Correlations**

| | | **MCQ1** | **MEQ1** |
|---|---|---|---|
| **MCQ1** | Pearson Correlation | 1 | 0.547(**) |
| | Sig. (2-tailed) | - | 0.000 |
| | N | 180 | 180 |
| **MEQ1** | Pearson Correlation | 0.547(**) | 1 |
| | Sig. (2-tailed) | 0.000 | - |
| | N | 180 | 180 |

** Correlation is significant at the 0.01 level (2-tailed).

**Correlations**

| | | **MCQ2** | **MEQ2** |
|---|---|---|---|
| **MCQ2** | Pearson Correlation | 1 | 0.691(**) |
| | Sig. (2-tailed) | - | 0.000 |
| | N | 180 | 180 |
| **MEQ2** | Pearson Correlation | 0.691(**) | 1 |
| | Sig. (2-tailed) | 0.000 | - |
| | N | 180 | 180 |

** Correlation is significant at the 0.01 level (2-tailed).

## REFERENCES

[1] Schuwirth LWT. General concerns about assessment. In: McCoubrie P, Ed. Improving the fairness of multiple-choice questions: a literature review. Med Teach 2004; 26: 709-12.

[2] Petrusa ER, Blackwell TA, Rogers LP, Saydjari C, Parcel S, Guckian JC. An objective measure of clinical performance. Am J Med 1987; 83: 34-42.

[3] Cureton EE. The upper and lower twenty-seven per cent rule. Psychometrika 1957; 22(3): 293-6.

[4] Lopez AT. The item discrimination index: Does it work? Rasch Measur Transact 1998; 12(1): 626.

[5] Man CC. Test reliability and item analysis with microcomputer. CUHK Educ J 1985; 13(2): 97-101.

[6] Buckley-Sharp MD, Harris FTC. Methods of analysis of multiple-choice examinations and questions. Br J Med Educ 1972; 6: 53-60.

[7] Linn RL, Gronlund NE. Measurement and assessment in teaching. 8th ed. New Jersey: Printice-Hall, Inc. 2000.

[8] Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. Med Educ 1988; 22: 325-34.

[9] Na-Tse TM. A life satisfaction questionnaire for Chinese women with schizophrenia. Hong Kong J Psychiatry 2003; 13: 7-16.

[10] Taylor R, Reeves B, Mears R, *et al*. Development and validation of a questionnaire to evaluate the effectiveness of evidence- based practice teaching. Med Educ 2001; 35: 544-7.

[11] Nnodim JO. Multiple-choice testing in anatomy. Med Educ 1992; 26: 301-9.

[12] Newble D. Assessment. In: Jolly B, Rees L, Eds. Medical Education in the Millennium. Oxford: Oxford University Press 1998.

[13] Brown B, Roberts J, Rankin J, Stevens B, Tompkins C, Patton D. The objective structured clinical examination: reliability and validity. In: Further developments in assessing clinical competence. (International conference proceedings) Ottawa, Canada 1987: pp. 563-71.

[14] Fraser R, McKinley RK, Mulholland H. Consultation competence in general practice: establishing the face validity of prioritized criteria in the Leicester assessment package. Br J General Pract 1994; 44: 109-13.

[15] Grand'Maison P, Lescop J, Rainsberry P, Brailovsky CA. Large-scale use of objective structured clinical examination for licensing family physicians. Can Med Assoc J 1992; 146: 1735-40.

[16] Dixon RA. Evaluating and improving multiple-choice papers: True-false questions in public health medicine. Med Educ 1994; 28: 400-8.

[17] Ebel RL, Frisbie DA. Essentials of educational measurement. 5th ed. New Jersey: Printice-Hall, Inc. 1991.

[18] Si-Mui S, Isaiah RR. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. Ann Acad Med (Singapore) 2006; 35(2): 67-71.

[19] Veloski JJ, Rabinowttz 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: 539-46.

[20] Brown B. Trends in assessment. In: Harden R, Hart I, Mulholland H, Eds. Approaches to the assessment of clinical competence. UK: Dundee Centre for Medical Education 1992; vol. 1.

Received: September 18, 2009 Revised: November 25, 2009 Accepted: December 20, 2009