Assessment of Family Doctors in Oman: getting the questions right Preliminary findings of a performance analysis of multiple choice questions

T Theodorsson, K El Shafie, N Al Wardy, A Khan, A Al Mahrezi, M Al Shafaee

Citation

T Theodorsson, K El Shafie, N Al Wardy, A Khan, A Al Mahrezi, M Al Shafaee. *Assessment of Family Doctors in Oman: getting the questions right Preliminary findings of a performance analysis of multiple choice questions*. The Internet Journal of Medical Education. 2009 Volume 1 Number 1.

Abstract

Background Properly constructed multiple-choice questions (MCQs) in high stakes examinations are expected to have high validity and reliability scores. However, several reports show that teacher-generated high stakes examinations do not always achieve the required high level of quality if item constructors are not trained in item writing, or if they are not proficient in the principles of assessment. AimThis evaluation aimed to assess the validity, reliability and quality of a 150 item multiple choice question test in the Membership of the Royal College of General Practitioners International Examination in Oman.
Design of the studyComputer-based test-item analysis according to a set of pre-validated quality criteria.
Participants and settingTwenty doctors who underwent Family Medicine Residency Programme of the Oman Medical Speciality Board, or its equivalent, and were eligible to sit the test.
Method The test-item analysis included item difficulty, item discrimination level and quality of distractors.
Results Across 150 A-type items, 69% were of applied format. Kuder- Richardson 20 was 0.81. The mean test score was 86.3% and standard error of measurement was

± 5.0. The mean difficulty index of the 150 items was 43%. Of all items, 50.7 % were at the level of moderate or better discrimination. Only 20% of items had more than two distractors functioning according to a quality criterion. Conclusion Distractor performance was found to be less than optimal and, if the time spent on test-item construction can be made more effective, that would be of great practical significance to teaching faculty. Despite the limitations of the study by low numbers of examinees, which impacts upon its validity, it is still the belief of the authors, that the analysis and suggestions made are useful as a guide to item writers, providing some answers as to how to improve the overall quality of MCQs in the future. To further improve this study it is now the intention to collect data from a larger number of subsequent examinations to increase the validity of the item analysis.

INTRODUCTION

In Miller's hierarchy of testing clinical competence and performance [1] testing of knowledge is the basic intention, followed by testing application of knowledge. A-type multiple choice questions (MCQ) of applied format and modified essay questions seek to test examinees' knowledge base and application of this knowledge in problem-solving, decision-making and management. Properly constructed MCQs in high stakes examinations are expected to have high validity and reliability scores [2]. However, several reports show that teacher-generated high stakes examinations do not always achieve the required high level of quality if item constructors are not trained in item writing, or if they are not proficient in the principles of assessment [3].

In addressing the issue of quality assurance in item writing, Ware and Vik [2] set out five quality criteria: i) strong adherence to a structured format, ii) the proportion of items of applied format shall be at least 50%, iii) of all distractors, 50% shall be functioning at 5% level, iv) at least 60% of items shall have moderate or better discrimination using set ranges, and v) the frequency of item-writing flaws agreed for the institution shall be less than 10%.

In 1993, the Department of Family Medicine & Public Health in the College of Medicine and Health Sciences Sultan Qaboos University in Oman developed a four year

residency programme in family medicine under the auspices of the Oman Medical Specialty Board (OMSB). The OMSB organizes and oversees all specialist residency programmes in Oman. In 1998, in collaboration with the Royal College of General Practitioners (RCGP-UK) the OMSB developed the Examination for Membership of the Royal College of General Practitioners International (MRCGP[INT]) with Oman as the first country to pilot this examination in 2001 [4].

In this article, we aim to evaluate the quality of the MCQ test which was one of the test-modules of the MRCGP[INT] examination that took place, in March 2009, in Oman. The purpose is to investigate the difficulty level of the test items and the quality of distractors, in particular.

METHODS

Twenty doctors who underwent the Family Medicine Residency Programme of the Oman Medical Speciality Board (OMSB), or its equivalent, sat a 150 item A-type MCQ test (a single best answer out of five options). The test was part of the endpoint assessment, and the items were, for the first time, constructed entirely by 12 senior faculty members who wrote the questions using the guidelines by the National Board of Medical Examiners in the USA [5]. They were not experts, but had been trained in a series of workshops on item-writing by the International Development Advisor from the RCGP-UK. The test was a pen-and-paper test of two and a half hours' duration. The test aimed to test core knowledge of medical practice in Oman with a main focus on clinical medicine, public health, evidence-based medicine and research methodology.

To improve the content validity of the MCQ test, the content of test-items was matched against the learning objectives and core topics of the curriculum of the Family Medicine Residency Programme. Each test-item focussed on a particular domain, such as diagnosis, investigation, drug treatment etc, and the subject category was selected from the core topics covered in the curriculum.

The 150 test-items were constructed according to a structured format, which was agreed upon by the group, and is depicted in Table 1. In the test paper itself, the examinees were given the theme, the stem, the lead-in question and the options to read. Emphasis was placed on writing MCQs with context-dependent test-items (i.e.with a clinical scenario testing application of knowledge and reasoning).

Figure 1

Table 1 Structured format of test-items when constructed

Category	The clinical system (e.g. cardiovascular for a theme of palpitations).
Domain	The domain tested (e.g. diagnosis for a theme of palpitations).
Theme	Allows signposting for candidates without giving away the answer (e.g. Palpitations or by default it is the category).
Stem	A clinical scenario with just enough information to be able to select the correct answer.
Lead-in-question	Very succinct and clear. Trying not to test negative knowledge. Trying to get candidates to integrate knowledge and make appropriate inferences from the information in the stem.
Options	Alphabetical order, short sentences, homogeneous (e.g. drugs, tests, etc based on the domain tested), avoiding negatives. Plausible but incorrect distractors.
Correct answer	Make sure it is correct in light of current knowledge
Background	Provide the evidence base and reference justifying the correct answer

A standard setting exercise was performed by a group of six senior faculty members, some of whom were the test-item writers, using the Angoff procedure augmented by the Hofstee procedure [6]. The passing score was set according to these at 50%. All 150 test-items were included in the marking.

The IDEAL-HK, Hong Kong item analysis software, version 4.0, was used to assess the performance of the 150 MCQ test- items [7]. The item analysis focused on reliability, item difficulty, discriminating power and distractor evaluation. The Kuder-Richardson 20 (KR-20) formula was used as a measure of reliability. A test-item with a difficulty index equal to or above 0.85 was set as being an easy item and a difficulty index equal to or below 0.20 as a difficult item [10].

Discrimination is another important concept for judging the quality of items [11]. We used the point-biserial correlation coefficient, as it is the most appropriate statistical procedure for correlation when one of the variables is a genuine dichotomy (which each item score is, i.e. correct or incorrect) [8]. We used a range reflecting three levels of discrimination power. A discrimination value of below + 0.19 indicates no significant discrimination power, whereas a value equal to or more than + 0.40 indicates excellent discrimination. Ware and Vik recommend that at least 60% of items should have moderate or better discrimination (i.e.> + 0.19) [2].

Distractors were evaluated according to how they were responded to. Various methods exist for evaluating distractor quality. In our analysis we used an evaluation based on

response frequency. Non-functioning or poorly performing distractors are usually defined as those that are chosen by less than 5% of examinees (2,3), but, since in our MCQ test the number of examinees was only 20, we chose to use equal to or less than 5%.

RESULTS

The item analysis (Table 2) showed that 104 items (69%) were constructed with a scenario testing applied knowledge, which is in line with the second quality criteria as suggested by Ware and Vik [2]. The Kuder-Richardson reliability coefficient (KR 20) of 0.81 (Table 2) indicates less than excellent reliability given the high stakes nature of our test [9]. The mean test score was 86.3 % and the standard error of measurement was ± 5.0.

In terms of difficulty, 30 items (20%) had a difficulty index (DI) of at least 0.85 or higher (easy, to too easy). Similarly, 20 items (13%) had DIs equal to or below 0.20 (very difficult). The average DI of the remaining 100 items was 0.55 compared with an average DI of 0.43 for all 150 items. As stated above, all 150 test-items were included in the marking.

Figure 2

Table 2: The general statistics of the item analysis

Variable	Number or %
Items	150
Examinees	20
Passing score	50 % (Angoff - Hofstee procedure)
Pass rate	85 %
Kuder-Richardson 20	0.81
Items testing application and reasoning	104 (69 %)
Items testing recall and comprehension	46 (31 %)
Test score mean	86.3 %
S.E. of measurement	5.09

In terms of discrimination, 76 of the items (50.7 %) were at the level of moderate or better discrimination (Table 3), and thus, well below the 60 % level of the fourth quality criterion recommended by Ware and Vik [2].

Figure 3

Table 3: Number of items in each discrimination category

Variable	Number
0.40 or above (very good discrimination)	26
0.30- 0.39 (good)	22
0.19- 0.29 (moderate)	28
< 0.19 (poor)	74

Regarding quality of distractors, (Table 4), of the 600 distractors only 284 distractors (47.3%) were functioning, thus not reaching the 50% level of the third quality criterion suggested by Ware and Vik [2]. Nineteen items (12.6%) had no functioning distractor and only 5 items (3.3%) had all four distractors functioning. Lastly, only 30 items (20%) had more than two distractors functioning (Table 5).

In addition to the results reported above, low-achieving examinees scored better than the high-achieving examinees in 33 items (22%). Of these, all but one had a discrimination index of less than + 0.19 and 14 of them (9.3%) had negative discrimination indices. Thus, of the 150 test-items, about one in five had a very low or non-existent discrimination value.

Figure 4

Table 4: Performance of distractors (out of 600 distractors)

Variable	Number (%)
Not selected	200 (33.3 %)
Selected by one examinee (i.e. 5%)	116 (19.3%)
Non-functioning	316 (52.6%)
Functioning	284 (47.3%)

Figure 5

Table 5: Functioning distractors per item (out of 150 items)

Number of distractors	Number of items (%)
Four	5 (3.3%)
Three	25 (16.6%)
Two	51 (34.0%)
One	50 (33.3%)
None	19 (12.6%)

DISCUSSION

SUMMARY OF THE MAIN FINDINGS

The main findings of our study are fourfold: Firstly, the proportion of test-items testing application of knowledge met the criterion set out by Ware and Vik [2]. Secondly, the average difficulty index of all 150 items in our examination was 43%, which is below a level of 60% regarded as the ideal for 5-option MCQs [10]. Thus, our MCQ test-items can be seen as having been rather difficult and that is reflected by a pass mark of 50% determined by Angoff and Hofstee procedure. Thirdly, the proportion of items with moderate to excellent discrimination power was 50.7 %, and thus short of the 60% criterion set by Ware and Vik [2]. Fourthly, only 20% of test-items had three or more functioning distractors.

STRENGTHS AND LIMITATIONS OF THE STUDY

These findings are based on the responses of the twenty examinees eligible to sit the test. The number of items (150) is not a problem from a psychometric point of view, but the low number of examinees is a problem and that poses a limitation to the validity of our conclusions. A solution to the limitation of our study would be to collect the results of a number of future examinations conducted along similar lines as our present study to enhance the validity of the analysis.

On the positive side, about 70% of the test-items were written in the applied format (with scenarios testing application of knowledge and reasoning), thus meeting the second quality criterion by Ware and Vik [2]. Furthermore, 20% of items were found to be easy and may be explained by well instructed, highly trained or highly able examinees. On the other hand 13% of items were difficult and may be explained by lesser able and less well trained examinees [11].

COMPARISON WITH EXISTING LITERATURE

The validity of the finding that only 20 % of our test- items had more than two functioning distractors is compromised by the low number of respondents in our study. However, as comparison, Tarrant et al. evaluated 541 items and found that only 13.8% had more than two functioning distractors [12].

These above-mentioned item-writing flaws are a cause for concern. On the other hand, item writers can expect that 50% or more of the items they write will fail to perform as expected [4]. Difficulties in designing plausible distractors are shared by most item writers constructing A-type MCQs with 4 or 5 options. Therefore, some researchers [11,12, 13] have argued that using MCQs of 3 option format would be just as reliable and valid from a psychometric point of view. Our distractor evaluation results might seem to lend strength to that idea, but again that result is inconclusive given the low number of examinees in our examination.

IMPLICATIONS FOR FUTURE ASSESSMENT AND RESEARCH

The 22% of test-items with discrimination indices of less than + 0.19 or negative, i.e. having very low or non-existent discrimination value, may indicate that those items were either mis-keyed (i.e. the option given, for the markers, as the key answer was not the correct one) or more likely intrinsically ambiguous [9]. This calls for better item construction, and perhaps, more training of item writers [3].

We would like to emphasize the importance of a thoughtful construction of plausible but still incorrect distractors and reinforcement of strong adherence to an agreed structured format of test-items. Evaluating distractor performance in teacher–generated tests is of interest, because the majority of tests that examinees take are teacher-generated and teachers spend a large amounts of time spent on test construction. If the time spent on construction can be made more effective, that would be of great practical significance to teaching faculty. It is only by carefully dissecting our assessment methods and content, and subjecting ourselves to test-item analysis that we can improve our system. The clarity and sound structure of MCQs is an increasingly important strategic concept in order to improve their validity [14]. There is no place for complacency if our assessment methods are to be used for international verification of competency.

HOW THIS FITS IN

Item writing guides emphasise that the validity of MCQs is enhanced by writing them in an applied format

Our study shows that the construction of MCQs of applied format is much improved by use of a structured format agreed upon by item writers.

Our study confirms the results of others in showing that constructing plausible but incorrect distractors is a difficult task.

Test-item analysis should include quality criteria to guide its interpretation in order to improve item construction, to facilitate decisions on which items to discard as too easy or too difficult and what distractors to replace.

Test-item analysis provides the necessary feedback to item writers to improve their question writing skills and is of no less an importance than proper blueprinting, content validity and item construction.

ACKNOWLEDGEMENTS

The authors are very grateful to, Prof. Raja C. Bandaranayake, International Consultant in Medical Education, Prof. Trevor Gibbs, Consultant in Medical Education & Primary Care,

Dr Adrian Freeman, International Development Advisor, MRCGP[INT] and Dr Ana Marusic, editor in chief Croatian Medical Journal, for kindly reviewing and providing most useful comments on this paper.

NOTES ON CONTRIBUTORS

Thord Theodorsson, MRCGP[INT], Senior Consultant, Dept of Family Medicine & Public Health, College of Medicine and Health Sciences, Sultan Qaboos University, Muscat Oman

Convenor of the MRCGP[INT] MCQ and written paper.

Kawther El Shafie, MRCGP[INT], Acting Consultant Dept

of Family Medicine & Public Health, College of Medicine and Health Sciences, Sultan Qaboos University, Muscat Oman

Co-convenor of the MRCGP[INT] MCQ and written paper.

Nadia Al Wardy, Assistant Professor, Head Medical Education Unit, Dept. of Biochemistry, College of Medicine and Health Sciences, Sultan Qaboos University, Muscat Oman.

Anwar Ali Khan, Associate Director, London Deanery GP Department, UK and International Development Advisor, MRCGP[INT] Oman.

Abdulaziz Al Mahrezi, MRCGP[INT], Senior Consultant, Dept of Family Medicine & Public Health, College of Medicine and Health Sciences, Sultan Qaboos University, Muscat Oman

Chairman of the Scientific Committee of the Family Medicine Residency Programme OMSB.

Mohammed Al Shafaee, MRCGP[INT], Assistant professor, Head Dept of Family Medicine & Public Health, College of Medicine and Health Sciences, Sultan Qaboos University, Muscat Oman. Chairman of MRCGP[INT] Examination Committee.

Ethics Committee

Medical Research and Ethics Committee, Sultan Qaboos University. The study does not require the Committee's approval as the primary purpose was as an evaluation.

References

1. Miller G E: The assessment of clinical skills/ competence/performance. Acad Med; 1990; 65 (Suppl): S63-67
2. Ware J, VIK T: Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. Medical Teacher; 2009; 31: 238-243
3. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew R: The quality of in-house medical school examinations. Acad Med; 2002; 77:156-161
4. Shafaee M: MRCGP[INT], the Omani Experience. RCGP International Newsletter; 2006; Issue 34, 1-2.
5. Case SM. Swanson DB: Constructing Written Test Questions for the Basic and Clinical Sciences, National Board of Medical Examiners 3rd ed. (revised); 2002. Downloaded from : http://www.nbme.org/publications/item-writing-manual.html
6. Bandaranayake R: Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Medical Teacher; 2008; 30:836-845
7. Precht C, Hazlett S, Yip J, et al: IDEAL Item Analysis User's Guide for Selected and Constructed Item Formats; 2nd ed.
8. Guilford JP, Fruchter B: Fundamental statistics in psychology and education; 6th ed. McGraw-Hill Book Company; 1978.
9. Hopkins KD: Educational and psychological measurement and evaluation; 8th ed. University of Colarado, Boulder; Allyn and Bacon; 1998.
10. The Division of Instructional Innovation and Assessment. Item analysis. Instructional assessment resources, The University of Texas Austin, accessed at: www.utexas.edu/academic/diia/assessment/iar/students/report/itemanalysis.php
11. Haladyna TM: Developing and validating multiple-choice test items; 2nd ed. Lawrence Erlbaum Associates; Publishers. Mahwah; New Jersey; London; 1999.
12. Tarrant M, Ware J, Mohammed AM: An assessment of functioning and non-functioning distracters in multiple-choice questions: a descriptive analysis. BMC Medical Education; 2009; 9:40
13. Vyas R, Supe A: Multiple choice questions: a literature review on the optimal number of options. Natl Med J India; 2008; May-June; 21(3):130-3.
14. McCoubrie P: Improving the fairness of multiple-choice questions: a literature review. Medical Teacher; 2004; 26 (8): 709-712

Author Information

Thord Theodorsson
Senior Consultant, Dept. of Family Medicine & Public Health, Sultan Qaboos University

Kawther El Shafie
Dept. of Family Medicine&Public Health, Sultan Qaboos University

Nadia Al Wardy
Medical Education Unit, Sultan Qaboos University

Anwar Khan
London Deanery, and Royal College of General Practitioners

Abdulaziz Al Mahrezi
Dept. of Family Medicine & Public Health, Sultan Qaboos University

Mohammed Al Shafaee
Dept. of Family Medicine & Public Health, Sultan Qaboos University