Multiple Choice Questions – An Introduction

Contents

Introduction

Multiple choice questions are a common method of assessment in the undergraduate assessment of medical students. Misconceptions still exist that only factual and recall of knowledge can be tested. Increasing evidence shows that clinical case based structured items can overcome the criticisms of the old multiple-choice questions and adherence to specified item construction can improve both the validity and reliability of MCQ's (Case and Swanson, 2001).

1. Multiple choice questions

MCQ test items are advantageous as they can sample large parts of a course and examine large numbers of students simultaneously. The tests can be used for both formative and summative assessments. Furthermore the use of computer based assessment promotes consistency in marking. Questions can also be banked for future use. Although MCQs do not allow for the evaluation of ideas, or the formulation of as arguments well-constructed multiple choice items are able to test all levels of students learning to encourage more than mere recall of theory.

To ensure the quality of questions are at a good standard the assessment should focus on testing relevant and core content. As the process to construct MCQs are time intensive, testing irrelevant content should be avoided. Different formats of MCQ test items are available:

- One best answer with 4 or 5 options with clinical or functional case scenario or vignette which includes gender, age, clinical signs and symptoms.
- Case cluster which is a 4 or 5 options MCQ with 4 items based on one clinical case presentation.
- True/false questions either 2 option or 4.
- Extended matching items (EMI) are one best answer MCQ with between 5 to 26 options using 2 or more items.
- Pick N items are similar to EMIs but have more than two correct options.

One best answer and case cluster questions are preferable as they can be developed to assess comprehension, application and problem solving skills. True/false test items should be avoided unless the answer is absolutely true or false (Chandratilake, et al., 2011).

Two option true/ false items increase the opportunity for guessing allowing students who guess to be correct 50% of the time; in the one best answer item with five options guessing is estimated at 20% (George, 2003). Pick N-items allow multiple correct options in cases where more than two items may be equally correct providing a better alternative to the true/ false items (Case and Swanson, 2001; Collins, 2006).

Although MCQs can be constructed as a direct question it is best to write the stem using a case scenario (or vignette), and then pose the question. The stem should have sufficient detail to allow the student to answer correctly without looking at the option list. Stems that consist of incomplete statements are not appropriate as they do not indicate clearly what the examinee needs to do.

To develop and improve the quality MCQ of questions it is important to consider the following:
- Be clear and explicit about the course objectives and the core competencies to test
- Ensure that the questions are aligned with the course objectives,
- Carefully determine the number of items needed for the test to meet these objectives
- Choose the best MCQ format for the test,
- Validate the questions and selected responses by providing references for the sources of information underpinning them. Referencing the questions documents their validity at a given time, and this is important if they may be used again in the future.

1.1 Extended matching items (EMIs)

EMIs combine the advantages of MCQs with the added advantage that they can test students' ability to solve problems rather than merely remember facts. The item set can test understanding by providing hypothetical scenarios and asking students to select most likely consequences. It can also present the students with a problem and possible solutions to encourage evaluation and synthesis of information. The students can then be asked to evaluate the solution based on criteria.

An EMI should have:
- a theme which indicates the subject that will be tested
- an option list with at least four distracters for the one correct option
- a lead-in statement which links the stem to the option list and tells the students what to do
- two or more item stems

Ideally the options list should have between 5 to 8 choices although this format can have up to 26 items; using more options increases the difficulty of the questions and reduces the chance of guessing (Bhakta, et al., 2005). The option list should be short statements or single words that do not inadvertently provide cues to test-wise students. (Swanson, et al., 2006).

The option list can be re used with different item sets (Beullens, et al., 2005; George, 2003). This saves time for constructing new items using the same option list. Diagrams, pictures and video clips can be used in the option list or as part of the stem. Using item sets with 2 or 3 stems allow for sampling content in a course more broadly whereas increasing the number of stems creates opportunity to test for the depth of content.

EMI sets in a MCQ test reduces the number of questions that would require if only standard one-best-answer 5 options format were to be used. In one best answer 5-option format, only 52 items can be administered in an hour compared to 54 with the 8-option EMI and 2 items sets (Case et al, 1994). When designing EMIs the number of options must be decided upon as using all 26 options increases the item difficulty but it also the time required to complete the test (Swanson, et al., 2006).

1.2 Pick N items

Pick N multiple choice items are similar to EMIs, but the item has two or more correct options. This ensures that this item do not become a one best answer or true and false item (Case and Swanson, 2001). This multiple choice type is well suited where there are multiple correct options and certain actions, or procedures can be done simultaneously or sequentially.

Including more EMI sets in a test decrease the number of standard one-best-answer 5 options format. In one best answer 5-option format, 52 items can be administered and completed by students in an hour compared to 54 with the 8-option EMI and 2 items sets (Case et al, 1994).

EMIs create opportunity to test more depth of content in one item where as the one best answer format single items need to be constructed for each level of learning. When designing EMIs the number of options must be decided upon as using all 26 options increases item difficulty but it also increase the time required to complete the test (Swanson, et al., 2006).

1.3 Writing item stems

The item is the question the examinee must answer; it should be unambiguously structured to elicit the correct option and not the wrong one.

Avoid using negative wording in the item. If negative wording is essential (e.g. "which is not a clinical symptom of COPD?", then the NOT should be in underlined, capitalised italics to ensure that the students attention is drawn to it.

The vignette is the stem or descriptive part of the question and should have all the detail to answer the item. Information that is not required in the stem should be removed. Patient vignettes or case scenarios are preferable to use as they improve the quality of questions.

1.4 Writing distractors

Distractors are used to discriminate between those students that know and those that do not know.
Write distractors that are:
a) accurate but do not meet the requirements of the problem.
b) Incorrect statements that may seem correct to the examinee.
c) Plausible/ likely but clearly incorrect.
d) Not clearly inappropriate.
e) Grammatically correct in relation to the stem.
f) If the stem is in the past tense all options should be in past tense; if the stem calls for plurals then the options should be plural. Distractors that flow grammatically from the stem are more likely to be chosen than those that do not, irrespective of the students' knowledge.
g) All options should be either in numerical (if the answer is a number) or alphabetical order. This makes it easier to read and review options and can reduce the time that the students spend on the question.

Do not use the following in any distractor:
a) None of the above or all of the above. This does not allow for selection of the best answer and only tests what the student recognises as either incorrect or correct.
b) Abbreviations, acronyms etc. without describing the term initially,
Distractors in one item should not allow the student to know the correct answer to another item in the test.
Avoid questions that expect the student to know the answer to the one question before being able to continue to the next question.

2. Psychometric testing

Basic Item analysis in MCQ is a numerical assessment of the test which includes item difficulty and discrimination. Poorly constructed and flawed items can make easy questions seem difficult because students who have the knowledge but do not understand the question give the wrong answer. Reviewing the test for formatting and accuracy, both internally and externally will improve the validity of the items. Validity relates to how well the test measures what it claims to measure in terms of the learning outcomes, including the level of learning to be tested and the appropriateness of the content.

Reliability refers to the generalization and reproducibility of the test. The score that the student obtains on a particular test should represent the score at any other given time and for similar tests (Luckett and Sutherland, 2000).

Item difficulty (p) indicates how easy or difficult the item is, and is determined by the number of correct responses to each item. An item that scores between 50-70% is considered to be moderately difficult; less than 50% is difficult, while above 75% is regarded as easy. An item with 70-90 % score indicates that the item is very easy as most of the examinees answered this correctly, while a score of less than 25% shows that the item is extremely difficult and may need to be revised (Phipps and Brackbill, 2009).

Item discrimination (D) refers to how well the assessment discriminates between high and low scoring students; the students who know the content and students who do not know. The discrimination index for each item in the test is calculated using the number of high scoring students who answered the item correctly and the number of low scoring students who answered the same item correctly (Zurawski, 1998). Commonly the upper 27 % and lower 27 % of students who have taken the test is used for the analysis provided the sample of test takers is large enough. The discrimination value should fall between -1 and +1 (McRowan and McRowan, 1999). The closer to + 1 the better the item discriminates between the responses of the 2 groups. This indicates that high performing students would select the correct answer for each item more often than the low scoring students. If this occurs then the assessment is said to have a positive discrimination index between 0 to +1. If the low scoring students selected the correct answer more often the assessment is said to have a negative index falling -1 to 0. An item with a negative index shows that poor students answered correctly more than good students and this item must be revised for item flaws. However if the item has a 0 discrimination it indicates that everyone either got the question right or wrong and that this question needs to be revised or removed as it is unable to discriminate between strong or weak students (Phipps and Brackbill, 2009).

Item analysis of the multiple choice items should not be looked at in isolation but taking into account the level of understanding tested and the year of study of the student.

3. Standard setting review process

Standard setting for a test involves agreement by experts in the course or subject to be tested on a pass or fail standard for the group of students based on the content, year of study or difficulty of the question (Norcini, 2003).There are various methods used in the standard setting process but the Angoff and Ebel Procedure are suggested for MCQs although the Hofstee method may also be used (Fowell et al, 2008). In the Modified Angoff method a panel of experts first discuss and agree on the characteristics of a borderline student. The panel then predict what percentage of the group would answer a given item correctly. The pass or fail standard is the average of the percentage correct response for the items.

In the Modified Ebel Procedure the relevance of each item in a test is rated based on the characteristics of the borderline student. A panel of experts rate each item's relevance using categories of essential, important or indicated. In each category the panel rate the expected number of borderline students that will answer the items correctly. Ideally consensus should be obtained between the participants when rating the items however an average of the panel's scores can also be calculated. The pass or fail standard for the test is determined by the percentage correct points of the borderline student (Case and Swanson, 2002).

4. Suggestions to assist development of quality MCQs

- Is the MCQ the appropriate assessment for the learning objective?
- Does the assessment reflect the course content?
- Is the grammar of the stem and options correct,
- Does the stem include all the relevant information?
- Are all options vertically below the stem for one best answer?
- EMIs should have a theme, options, lead in statement and item stem.

- All options are equal in length, style, format and grammar
- Is one option clearly correct?

References

- Bauer, D., Holzer, M., Kopp, V. and Fischer Martin R. 2011. Pick-N multiple choice-exams: a comparison of scoring algorithms. Advances in Health Sciences Education: Theory Practice. Abstract 16(2):211-21.

- Beullens, J., Struyf, E., Damme, B. V. 2005. Do extended matching items multiple choice measure clinical reasoning. Medical Education, 39 (4) 410-417

- Case and Swanson. 2001. Constructing Written Test Questions for the Basic and Clinical Sciences. Third Edition (Revised)

- Case, S.M., Swanson, D.B, and Ripkey, D.R. 1994. Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. Academic Medicine, Abstract 69 (10 Suppl):S1-3.

- Chandratilake, M, Davis, M., Ponnamperuma, G. 2011. Assessment of medical knowledge: the pros and cons of using true/false multiple choice questions. The National Medical Journal of India. Abstract; 24(4):225

- Collins, J. 2006. Writing Multiple-Choice Questions for Continuing Medical Education Activities and Self-Assessment Modules. RadioGraphics; 26:543–551

- Fowell, S.L., Fewtrell, R., McLaughlin, P.J. 2008. Estimating the minimum number of judges required for test-centred standard setting on written assessments. Do discussion and iteration have an influence? Advances in Health Science Education Theory Practice. 13(1):11-24.

- George, S. 2003. Extended matching items (EMIs): solving the conundrum. The British Journal of Psychiatry, (27) 230-232

- Luckett, K and Sutherland L. 2000. Assessment practices that improves Teaching and Learning, Experiencing a Paradigm shift Through Assessment, Chapter 1, Huba and Jann

- Norcini, J. 2003 Setting standards on educational tests. Medical Education, 37:464–469

- McCowan R and McCowan S. 1999. Item Analysis for Criterion-Referenced Tests. Available at: http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED501716&ERICExtSearch_SearchType_0=no&accno=ED501716. Accessed 9 January 2012.

- Phipps, S.D and Brackbill, L. 2009. Relationship Between Assessment Item Format and Item Performance Characteristics. American Journal of Pharmaceutical Education 73(8): 146.

- Ripkey, D.R., Case, S.M and Swanson, D.B. 1996. A "New" Item format for Assessing Aspects of Clinical competence. Academic Medicine, 71 (10) 34-37.

- Swanson, D.B., Holtzman, K.Z., Allbee, K. 2008. Measurement characteristics of content-parallel single-best-answer and extended-matching questions in relation to number and source of options. Journal of the Association of American Colleges, 83(10 Suppl):S21-4

- Swanson, D.B., Holtzman, K.Z., Allbee, K., Clauser, and B.E. 2006. Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of

options. Journal of the Association of American Colleges. Abstract 81(10 Suppl):S52-5

- Zurawsk, R. M.1998. Making the Most of Exams: Procedures for Item Analysis. The National Teaching and Learning Forum. (7) 6