Medical Education

Assessment of medical knowledge: The pros and cons of using true/false multiple choice questions

MADAWA CHANDRATILAKE, MARGERY DAVIS, GOMINDA PONNAMPERUMA

ABSTRACT

True/false multiple choice items, commonly referred to as true/false multiple choice questions (MCQs), were previously a widely used selected response examination format. They can be written relatively easily and cover a wide range of content. Educational researchers have however highlighted several adverse features of this format that make it inappropriate for many assessment settings. These include: (i) there is a high chance of guessing the correct answer; (ii) some marks are not awarded for knowing the correct answer, but for knowing that an answer is incorrect; (iii) they are weak in discriminating between high and low performers; (iv) identifying items which are absolutely true or false may lead to assessment of trivial knowledge; (v) there are difficulties with constructing flawless items; (vi) they may not encourage learning around the items; and (vii) they may not assess what they purport to assess. Many assessment agencies abandoned the use of this format decades ago due to these shortcomings.

The use of single best answer (SBA) and extended matching item (EMI) formats helps overcome or minimize the above weaknesses. Assessors who plan to change to SBA or EMI formats from true/false MCQs may, however, need to increase the number of questions to include a representative sample of the curriculum (lengthening the question paper). However, they may not need to increase the examination time, as in general students can answer more SBAs or EMIs than true/false MCQs per unit time.

It is time that we reflect upon the disadvantages of true/false MCQs and review their place in our assessment toolkit, as their use in summative examinations may not be fair to students, especially 'good' students.

Natl Med J India 2011;24:225–8

INTRODUCTION

Selected response questions (e.g. multiple choice questions [MCQs]) are a popular method of assessing medical knowledge all over the world.[1] The advantages of this method of assessment include the potential for high validity and reliability, ease of marking answer scripts, usefulness as an aid to learning and self-assessment, and high objectivity.[2]

The true/false multiple choice item, commonly referred to as true/false MCQs, was a widely used selected response format.[3] In the assessment of knowledge, true/false MCQs can be used to cover a large content area and can be written relatively easily, especially at the levels of factual recall and comprehension.[3] However, in the late 1980s, educational researchers highlighted several limitations of this format that make it inappropriate for many assessment settings.[4] As a result, it lost its place in the summative assessment toolkits of many assessment agencies, especially in the UK and North America. However, in some countries, especially in Asia, true/false MCQs are still being used summatively both in undergraduate and postgraduate medical education. We discuss several key 'concerns' related to the use of the true/false format and propose alternative formats, which lead to better educational outcomes.

The shortcomings of true/false MCQs include the following: (i) there is a high chance of guessing the correct answer;[4] (ii) marks are not awarded for knowing the correct answer, but for knowing that an answer is incorrect;[1] (iii) they are weak in discriminating between high and low performers;[4] (iv) identifying items which are absolutely true or false may lead to assessment of trivial knowledge;[4,5] (v) there are difficulties with constructing flawless items;[1,6] (vi) they may not encourage learning around the items;[7] and (vii) they may not assess what they purport to assess.[8]

SHORTCOMINGS OF TRUE/FALSE MCQs

High chance of guessing the answer

Since there are only 2 options (true or false) from which the candidate has to choose the correct one, each option has a 50% chance of being selected as the right answer even if the candidate is completely ignorant about the test material. Therefore, the candidate has a 50% chance of correctly answering each question item by pure guess work. Guessing reduces the reliability of assessment results.[2] In the past, many assessors attempted to counteract the effects of guessing on reliability by awarding negative marks for incorrect responses to items, i.e. negative marking. However, negative marking does not significantly improve reliability.[9] Awarding negative marks for incorrect responses makes matters worse for the following reasons:

1. Reducing marks for incorrectly answered items is a penalty.[10] Though rewards reinforce the desired behaviour permanently or at least in the long term, punishments discourage undesired behaviours only temporarily.[11] Therefore, negative marks may not necessarily prevent students from guessing.
2. Assessment marks usually provide feedback to students, which is essential for them to reflect upon their performance.[1,5] However, with negative marking, the feedback received by the students is not specific, but confusing. Therefore, the students

University of Dundee, Tay Park House, 484 Perth Road,
Dundee DD2 1LR, Scotland, UK

MADAWA CHANDRATILAKE, MARGERY DAVIS
Centre for Medical Education

University of Colombo, Sri Lanka

GOMINDA PONNAMPERUMA Faculty of Medicine

Correspondence to MADAWA CHANDRATILAKE;
m.chandratilake@dundee.ac.uk

cannot evaluate their knowledge or ignorance and plan for future learning. Even the assessors cannot clearly and confidently understand the students' level of knowledge by reviewing their marks.

3. Negative marking may also reward high risk-takers. For example, if a student scores 75% in a 100-item examination that awards negative marks, the 75% score can be a result of several score combinations depending on the number of questions he/she guessed, partially guessed or left unanswered. Proponents of negative marking may say that students should not have guessed or partially guessed. However, in reality it is very difficult, if not impossible, to prevent guessing— especially partial guessing (guessing the answers only for some items) or educated guessing (guessing based on experience, knowledge or other information without knowing the correct answer). If this candidate was a person who followed the examiner's instructions to the letter, he may have attempted only 75 items. Another candidate who had the same amount of knowledge but was a higher risk taker could score 80% by attempting 90 items, out of which 75 items he/she knew for sure were correct, and correctly answered another 10 of the remaining 15 items with educated guessing. Of the two candidates with the same amount of knowledge, one scores more in the same examination because he/she is a higher risk-taker than the other. Therefore, this examination has rewarded not only knowledge, but also risk-taking. The tendency for risk-taking differs significantly with gender[12,13] and personality.[12] Therefore, by asking students not to guess or by punishing them for guessing, the examiners cannot prevent or reduce guessing as long as personality and gender differences among students exist; some are risk-takers and some are not. What the examiners can ensure, however, is not to unduly reward the personality differences of students. This can be achieved by refraining from awarding negative marks.

Marks not awarded for knowing the correct answer

In the true/false format, a student who indicates an incorrect statement as false receives marks. He/she may have indicated it as false, thinking another incorrect answer was the correct answer. Therefore, the student might not have known the correct answer but is still awarded marks.[4,14]

Weak in discriminating high and low performers

In medical education, students who achieve the required standards may be ranked in order to award classes and/or distinctions. However, Oosterhof and Glasnapp[15] and Ebel,[16] all of whom examined discrimination indices of questions, concluded that true/false questions have less power than single best answer (SBA) questions to discriminate between high and low performers. Negative marking may further reduce this ability.[10]

Absolute true or false items may lead to the assessment of trivial knowledge

In practice, the solutions for many medical problems are neither 'black' nor 'white', but rather, 'the best for the situation'. Therefore, in the context of medicine, some correct answers are not always correct and some incorrect answers are not always incorrect.[1] The attempt to identify items that are absolutely true or false may lead to the assessment of trivial knowledge. True/false questions, therefore, may restrict the examiners' ability to assess important content areas that cannot be reduced to either true or false statements.

Difficult to test higher order thinking

Memorizing facts, which requires lower order thinking is important in medicine. As facts can only be either correct or incorrect, true/false questions may be suitable to assess recall of factual information.[4,17,18] However, developing the ability of students to solve clinical problems is one of the main goals of medical education,[19] and imparting factual knowledge is only one aspect of this goal.[7] Usually, higher order thinking is assessed in MCQs by constructing context-rich questions, i.e. questions based on clinical or practical scenarios.[7] For the reasons described above, arriving at an absolute correct or incorrect answer to a given case scenario may be difficult.[20] However, it would be appropriate to ask for a best possible answer in such a context and this would be more acceptable than asking for an absolutely correct or incorrect answer. Therefore, most assessors will find it easier to construct context-rich questions with SBA and extended matching formats than with the true/false format, allowing candidates to express their judgement as to the best option for a specific clinical situation.

Difficult to write flawlessly

True/false items are more likely to have identifiable technical flaws compared to other selected response formats.[1,6] Technical flaws lead to the testing of the test-wiseness of students rather than their knowledge and cause irrelevant difficulty.[5] Some of the common flaws associated with true/false MCQs are given in Table I.

Studies of the technical flaws of true/false items have shown that these flaws increase examinees' scores and may help poor examinees at the expense of better ones.[6,21]

May not encourage learning around the items

The true/false MCQ format usually tests recall of factual knowledge, promoting lower order thinking. As learning is driven by the type of assessment,[7] students who are tested by this format of assessment attempt to memorize items of knowledge and may not be interested in learning deeply around the subject. Rees[22] showed that the medical students of Guy's Medical School in London did not learn areas related to the items in general, even though feedback (the correct answer and clarification) was provided for each true/false item. However, by using some of the other selected response formats, higher order thinking, such as the examinee's judgement about the best option for the situation, can be encouraged relatively easily, providing a focus for deep and around-the-topic learning.[4]

May not assess what it purports to assess

A valid assessment instrument should measure what it purports to measure.[8] For all the reasons discussed above, true/false MCQs do not measure exactly what is intended, thus weakening the validity of overall assessment.

ALTERNATIVE FORMATS

The use of true/false MCQs is not justified, therefore, in summative assessment of medical knowledge. Many assessment agencies, such as the National Board of Medical Examiners in the the USA and the Joint Committee on Intercollegiate Examinations for Surgery in the UK, no longer recommend this format of selected response questions. Rather, they recommend the use of the SBA[4,6,14] and extended matching item (EMI)[1,23] formats, which have been shown to be superior alternatives to the true/false MCQ format. Examples of SBA and EMI formats are given in Tables II and III.

TABLE I. Common flaws with true/false multiple choice items

Use absolute, frequency and vague terms to make the items absolutely true or false

— *Use of absolute terms (e.g. never)*

In haemolytic anaemia

a) spherocytosis is a recognized feature in a blood film under microscopy
b) normal haemoglobin levels in blood analysis can ***never*** be expected
c) skull and skeletal deformities can occur in infancy and early childhood
d) blood transfusion therapy may be considered for patients with severely compromised cardiopulmonary status

Option *b* contains an absolute term, unlike options *a, c* and *d.* The test-wise students will recognize that option *b* is less likely to be a true option than *a, c* and *d.*

— *Use of vague and frequency terms (e.g. usually, often, rarely)*

A patient who developed myocardial infarction

a) **often** shows no clinical features
b) **usually** complains of chest pain across the anterior precordium
c) **rarely** present with syncope
d) should be immediately treated with aspirin

Research has shown that consensus on the meaning of terms such as *usually*, *often* and *rarely* is difficult to achieve with content experts themselves.[20]

Use of phrases rather than questions

Regarding asthma

a) Asthma is the most common chronic disease in childhood
b) Exercise-induced asthma is primarily seen in persons with pre-existing asthma
c) Acute bronchoconstriction is caused by immunoglobulin E-dependent mediator
d) Predominantly occurs in boys in childhood

Some MCQ writers write *stems* of the MCQs as phrases rather than questions to cover a larger content area, as the items assess individual facts related to a topic instead of a concept as a whole.

This MCQ does not pose a clear question and it is not adequately focused. Examinees might have to go through the *stem* and the *items* over and over again to make sense of the items, making the question unnecessarily difficult.

TABLE II. Example of a single best asnwer (SBA) question

David is 6 months old. His mother brings him to the clinic. She is concerned that he does not look at her, does not speak and is not apparently interested in her or other members of his family.

What is the most likely diagnosis from the list below?

a) Congenital blindness*
b) Autism
c) Severe learning disability
d) Attachment disorder
e) Normal development

* correct option

TABLE III. Example of extended matching item (EMI) questions*

Urinary tract infection

a. Antegrade pyelogram
b. CT scan
c. DMSA scan
d. DTPA
e. Helical CT
f. Hippuran renogram
g. Isotope GFR
h. IVU
i. KUB
j. MAG 3scan
k. MCUG
l. MRI
m. MRU
n. Retrograde pyelogram
o. Retrograde urethrogram

For each of the following scenarios which is the single most appropriate investigation to perform?

1. A 2-year-old girl who has had a febrile illness and a proven urinary tract infection on two prior occasions has been diagnosed once again with pyelonephritis. (c)
2. A 7-year-old with a straddle injury presents with retention of urine, gross scrotal swelling and perineal bruising. (o)
3. A 7-year-old with known moderate hypertension presents with continuous wetting from birth. (m)

* Correct answers are given within brackets.

TABLE IV. Comparison of true/false multiple choice questions with single best answer (SBA) and extended matching item (EMI) questions

	True/False	Single best answer	Extended matching item
Chance of guessing	50%	20% (if 5 options)	12.5% (if 8 options)
Marks awarded for knowing the correct answer	Not always	Yes	Yes
Discriminating between high and low performers*	Poor	Good	Good
Level of knowledge assessed	Factual recall. Some of the facts may be trivial	Higher order thinking plus factual recall	Higher order thinking plus factual recall
Difficulty in writing flawlessly	High	Variable	Variable
Encouraging learning around the items*	No	Yes	Yes
Assessing what it purports to assess (i.e. validity of the assessment instrument)	Low (even if the content is adequately sampled)	High (if the content is adequately sampled)	High (if the content is adequately sampled)

*Based on research findings

ADVANTAGES OF ALTERNATIVE FORMATS

The use of the SBA and EMI formats helps to overcome or minimize the weaknesses identified in Table IV. In the SBA and EMI formats, the possibility of guessing is considerably lower than in the true/false format. In the commonly used SBA type, the students are expected to choose the best answer out of 5 options.[4] Therefore, the chance of guessing the correct answer is 20%. In an EMI, the options list ranges from 3 to 26.[5] The higher the number of options in the EMI option list, the lower the chance of guessing. Research has shown that acceptably reliable results with lower testing time can be achieved with EMIs with 8 options.[24] This is a practicable number of options in writing EMIs, with the chance of guessing the correct answer for each question being 12.5%. Therefore, items need not be subjected to negative marking, students are rewarded for their knowledge, they receive good feedback as the marks correspond to what they know; and assessors are confident about the students' achievement of standards by reviewing their marks. Both the SBA and EMI formats can be used effectively to assess higher cognitive levels. The SBA format is also cost-effective in assessing factual knowledge. Because of the above strengths, SBA and EMI assessment results should be more reproducible (i.e. high in reliability) than the results of a true/false assessment, if the questions are written appropriately and in adequate numbers.

Both at the undergraduate and postgraduate levels, EMIs have outperformed true/false MCQs in discriminating between high and low performers.[25] The discrimination ability of SBA questions has also been shown to be appreciable.[26] However, the EMI format seems to be the best option in terms of discriminating between high and low performers.[26] Both the SBA and EMI formats transform the 'black' or 'white' nature of the answers expected by true/false questions to the 'most appropriate answers for given situations'. If the SBA questions and EMIs are written correctly, without technical flaws, and focus on higher levels of knowledge, they can be used more effectively to assess problem-solving ability rather than factual recall and concepts.[5]

Sampling

Adequate sampling of the curriculum is an essential prerequisite for content validity in the case of any assessment method.[5] A large number of questions may be necessary to achieve adequate sampling, e.g. many SBA examinations comprise 175–250 questions.[27] Therefore, an increase in the number of items may be necessary when changing from the true/false to the SBA format. However, there may not be a need to increase the number of questions in the examination by too much as long as the examination includes a representative sample of the curriculum and produces reliable results.

Length of examination

It is assumed that the time allotted for SBA MCQ examinations is a function of the number of questions and options.[28] Although the number of SBA questions capable of being assessed per unit time may be less than the number of true/false questions, there may be no difference in the amount of information elicited by the 2 question formats. The reason for this is that in true/false MCQs, approximately half the marks come from false statements. The examiners do not know whether the candidates know the correct answers for these false statements. Therefore, the testing time corresponds to candidates' 'true' knowledge and becomes nearly half. For example, if there are 40 true/false questions in a 40-minute paper, and half of them require the answers to be 'false', only 20 questions will assess candidates' 'true' knowledge; therefore, the actual testing time corresponds to 20 minutes. In contrast, each SBA question elicits the candidates' 'true' knowledge. Therefore, the number of SBA questions included in a paper which is allocated 40 minutes can be approximately doubled compared to true/false MCQs. Increasing the number of SBA questions, therefore, does not necessarily lengthen the testing time used by existing true/false examinations. The rule of thumb is one item per minute for SBA MCQs.[24]

However, though true/false MCQs are unsuitable for summative assessments, this format can be effectively used for formative purposes, especially in quizzes, where students' understanding of teaching can be assessed after teaching sessions. The format allows for testing of a wide spread of issues within one item, giving feedback to both teacher and student, as in the first example in Table I.

Writing SBA and EMI formats of MCQs needs training. The question developers should be aware of the structure and the technical aspects of these formats based on the best evidence in medical education. Such evidence and guidance can be readily accessed online (*http://www.nbme.org*).

CONCLUSION

The true/false MCQ format of selected response questions is relatively easy to write and can be used to cover a large content area with fewer questions than other formats such as SBA and EMI. However, evidence in medical education has demonstrated that the true/false MCQ format is less valid and less reliable, and has a poorer educational impact than the other selected response formats in the assessment of medical knowledge (Table IV). Negative marking does not improve its validity and reliability. Many assessment agencies have abandoned its use decades ago due to these shortcomings. It is time we review the place of true/false MCQs in our assessment toolkit.

REFERENCES

1 Schuwirth LW, van der Vleuten CP. ABC of learning and teaching in medicine: Written assessment. *BMJ* 2003;**326:**643–5.

2 Nnodim JO. Multiple-choice testing in anatomy. *Med Educ* 1992;**26:**301–9.

3 Anderson J. Multiple choice questions revisited. *Med Teach* 2004;**26:**110–13.

4 Downing Steven M. True-false, alternate-choice, and multiple-choice items. *Educ Meas: Issues and practice* 1992;**11:**27–30.

5 Case SM, Swanson DB. *Constructing written test questions for the basic and clinical sciences.* 3rd ed (revised). Philadelphia:National Board of Medical Examiners; 2002. Available at *http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf* (accessed on 14 Jul 2011).

6 Albanese MA. Type K and other complex multiple-choice items: An analysis of research and item properties. *Educ Meas: Issues and practice* 1993;**12:**28–33.

7 Schuwirth LWT, van der Vleuten CPM. Different written assessment methods: What can be said about their strengths and weaknesses? *Med Educ* 2004;**38:**974–9.

8 Bridge D, Musial J, Frank R, Roe T, Sawilowsky S. Measurement practices: Methods for developing content valid student examinations. *Med Teach* 2003;**25:** 414–21.

9 Burton RF. Misinformation, partial knowledge and guessing in true/false tests. *Med Educ* 2002;**36:**805–11.

10 What is the right choice? Available at *http://www.timeshighereducation.co.uk/* (accessed on 7 Jul 2009).

11 Gage NL, Berliner David C. Operant conditioning: A practical theory. In: *Educational psychology.* 2nd ed. Chicago:Rand McNally College Publishing Company; 1979: 277–8.

12 Pawlowski B, Rajinder A, Dunbar RIM. Sex differences in everyday risk-taking behavior in humans. *Evol Psychol* 2008;**6:**29–42.

13 Gullone E, Moore S. Adolescent risk-taking and the five-factor model of personality. *J Adolesc* 2000;**23:**393–407.

14 Frisbie DA. Multiple choice versus true-false: A comparison of reliabilities and concurrent validities. *J Educ Meas* 1973;**10:**297–304.

15 Oosterhof AC, Glasnapp DR. Comparative reliabilities and difficulties of the multiple-choice and true-false formats. *J Experimental Educ* 1974;**42:**62–4.

16 Ebel RL. Are true–false items useful? In: Ebel RL (ed). *Practical problems in educational measurement.* Lexington, MA:D.C. Heath; 1980:145–56.

17 Scouller K, Prosser M. Students' experiences in studying for multiple choice question examinations. *Studies Higher Educ* 1994;**19:**267–79.

18 Downing SM, Baranowski RA, Grosso LJ, Norcini JJ. Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. *Applied Meas Educ* 1995;**8:**187–97.

19 Spencer JA, Jordan RK. Learner centred approaches in medical education. *BMJ* 1999;**318:**1280–3.

20 Case SM. The use of imprecise terms in examination questions: How frequent is frequently? *Acad Med* 1994;**69** (10 Suppl):S4–S6.

21 Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Med Educ* 2008;**42:**198–206.

22 Rees PJ. Do medical students learn from multiple choice examinations? *Med Educ* 1986;**20:**123–5.

23 Fenderson BA, Damjanov I, Robeson MR, Veloski JJ, Rubin E. The virtues of extended matching and uncued tests as alternatives to multiple choice questions. *Hum Pathol* 1997;**28:**526–32.

24 Swanson DB, Holtzman KZ, Allbee K, Clauser BE. Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. *Acad Med* 2006;**81** (10 Suppl):S52–S55.

25 Duthie S, Fiander A, Hodges P. EMQs: A new component of the MRCOG Part 1 examination. *Obstet Gynaecol* 2007;**9:**189–94.

26 Swanson DB, Holtzman KZ, Clauser BE, Sawhill AJ. Psychometric characteristics and response times for one-best-answer questions in relation to number and source of options. *Acad Med* 2005;**80** (10 Suppl):S93–S96.

27 Toolbox of assessment methods. Available at *http://www.acgme.org/outcome/assess/toolbox.asp* (accessed on 3 Sep 2008).

28 Case SM, Swanson DB, Ripkey DR. Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. *Acad Med* 1994;**69** (10 Suppl):S1–S3.