Fairness in using negative marking for assessing true/false questions[1][2]

Firoza Haffejee - Durban University of Technology, South Africa
Thomas E. Sommerville - University of KwaZulu-Natal, South Africa

ABSTRACT

Multiple choice questions are a popular method of testing as they are easy to mark. However, they also lend themselves to guessing. True/False questions are particularly prone to guessing. This can be alleviated by deduction of marks for incorrect answers. The University of KwaZulu-Natal currently uses true/false questions in its assessments, which are answered by means of pencil-and-paper sheets that are optically scanned, marks being calculated according to a standard formula. A trial of a proposed new computer program revealed mark discrepancies compared to the current scheme. This study evaluates the program's marking scheme 'internal negative marking' – whereby an overall negative mark for an item consisting of a stem and several true/false questions is 'rounded up' to zero. We enumerate the cause of the discrepancies, demonstrate that the latter scheme diminishes the penalty for guessing and may encourage strategic students to leave out parts of the curriculum when studying.

INTRODUCTION

It is claimed that multiple choice questions (MCQs) avoid the lack of reliability that is evident in the marking of essays and short answer questions (Hammond et al., 1998). Furthermore MCQ tests can assess a broad spectrum of the syllabus in a relatively short period of time (Moss, 2001; Brady, 2005; Manogue et al., 2002). Many assessments – particularly in science-based disciplines – are thus limited to MCQs (Davies, 2000). Despite ongoing questions about the extent to which MCQs conform to the requirements of a good assessment instrument (Davies, 2000; Moss, 2001; Burton, 2005; Ibrieger, 2006; McCoubrie, 2004), there is general agreement that they are a practical means of assessing the increasing numbers of learners with which higher education is faced (Hannan, English and Silver, 1999). Concomitant with this increase in numbers, the use of computers for assessment has steadily been increasing (Akdemir and Oguz, 2008). MCQ tests are easy to score automatically (Davies, 2000; Morrison and Free, 2001; Moss, 2001; Brady, 2005) and computers can instantly score the answers (Bugbee and Alan, 1996; Davies, 2000; Manogue et al., 2002). Computerisation in turn brings its own challenges (Zakrzewski and Steven, 2003) including 'pedagogic, operational, technical and financial issues'. However, it does allow automation of interventions such as negative marking introduced to mitigate the effect of guessing.

[1] Date of Submission 11 January 2014
Date of Acceptance 9 June 2014

[2] The authors are grateful to Ms N.M. Cetywayo for assisting with data entry and to Dr J.M. Van Wyk for critically reviewing the manuscript.

Negative marking

Virtually all forms of MCQs encourage guessing (Collins, 2006; McHarg et al., 2005). Random answering can result in some questions being answered correctly by chance alone (Carneson, Delpierre and Masters, 1996). In true/false questions – which can be regarded as MCQs with only two options – there is a 50% chance of guessing the correct option. If the passing score is set at 50%, a pass can therefore be obtained without any knowledge of the subject (Bandaranayake, Payne and White, 1999; Smoline, 2008). Random guessing (as distinct from informed or calculated answering) can be discouraged by deduction of marks for incorrect answers, which improves test reliability (Burton, 2004; McHarg et al., 2005). If negative marking is to achieve its goal – to discriminate between different students' performance – then the penalty applied should be great enough to discourage random guessing (Holsgrove, 1992) and to keep the span of meaningful marks as wide as possible. This is referred to as 'reliability length' (Burton, 2004). The formula generally used to calculate the quantum of negative marking is: I = C/ (n-1) where I = mark deducted for each incorrect alternative (distractor), C = mark awarded for the correct response, n = total number of alternatives and n-1 = the total number of incorrect alternatives (Carneson, Delpierre and Masters, 1996; Holt, 2006; Burton, 2004). Thus for a best-of four MCQ in which the mark for a correct answer is 1, choosing one of the three distractors would incur a mark of – ⅓. For a true/false question, the formula would yield a mark for an incorrect answer of –1 in the case where the correct answer scores +1.

Alternatives to negative marking

Although the negative marking formula is well-known and makes mathematical sense, it must be conceded that its use is not unquestioned. It has been questioned whether negative marking serves pre-emptively to discourage examinees from guessing, or retroactively to correct their mark for guesses made during testing, or in fact relates to an entirely different construct – examinees' risk-taking propensities (Burton, 2005; Goldik, 2008; Fowell and Jolly, 2000; Bernardo, 1998; Betts et al., 2009). For these and other reasons, the standard formula is not the only way of attempting to accommodate examinees' guesses. For example, the application of the formula can be such that within a group of true/false statements relating to a stem, a multitude of negative marks can be rounded up to zero; Cook (2010) refers to this as 'internal negative marking'. The marking scheme can be adjusted to allow examinees to indicate with what degree of confidence they answer each question, the mark allocated for a correct or incorrect answer then varying according to that confidence rating (Bauer et al., 2011; Cisar et al., 2009). Marks can be allocated according to combinations of answers selected or omitted (Chang et al., 2007; Jennings and Bush, 2006). Formulae can be applied that take into account the number of items in the test, the number of choices within each item, and examinees' mean scores and their variances (Zimmerman and Williams, 2003). Finally, the pass mark can be altered to take into account the likely score achieved by random guessing.

Localisation of a problem

The Nelson R Mandela School of Medicine at the University of KwaZulu-Natal uses a combination of true/false, MCQ and short-answer responses in tests and examinations in its undergraduate curriculum. These assessments are currently taken as a paper-based version where the true/false and MCQ answers are recorded on a sheet that can be scanned, and the standard negative marking formula applied. The faculty has, however, been exploring the option of changing over to wholly computer-based assessments, using the same format as in the paper-and-pencil version (Bugbee and Alan, 1996). Irrespective of whether responses are paper-based or captured directly on computer, the score of the assessment should be identical (Peak, 2005; Akdemir and Oguz, 2008). On a test run of the computer program under consideration, the researchers noted discrepancies in the results of true/false questions and set out to explore these further, in order to understand the discrepancies and their implications.

MATERIALS AND METHODS

An assessment which had already been taken by a group of students was used in the comparison between marking by scanning a paper sheet and by the direct response program. The assessment comprised a total of 150 true/false statements grouped into 34 items. Most of the items had five statements, although some had three, four or six. The assessment had been previously answered by the students in pencil on an MCQ answer sheet. One mark was allocated for each correct answer given and one mark was deducted for each incorrect answer. The scripts were scanned and marked electronically and a sample of the scripts was double-checked manually, according to the standard practice in the School of Medicine.

Nine of these scripts which had already been assessed were chosen by stratified random selection: three scripts each were taken from the highest, middle and lowest performance of the class. The student details were removed from the answer scripts to render them anonymous. The answers on the scripts were then re-marked and scores re-calculated using the computer program under consideration; one investigator (FH) acted as 'student', entering the nine students' original responses directly into the new program. All entries were double-checked. The marks obtained from the computer program after this re-assessment were then compared to those obtained in the paper-based scanned assessment. The scripts were also marked manually.

Since all the computer-based test scores differed from their paper-based counterparts, the following mock scripts were created in order to test our hypothesis that the discrepancies were due to the way in which the MCQ scanning program and the computer-based test program totalled the scores. Two scripts were created, each with 75 correct answers and 75 incorrect answers. In script A, the answers were distributed such that all the correct answers were clustered in the first half and all the incorrect answers in the latter half of the paper. In script B, the correct answers were evenly distributed throughout the paper – for example, in items with five statements, half of these items had three correct and two incorrect answers and the other half had two correct and three incorrect answers. Again, 75 options were answered correctly and the other 75 incorrectly. The same investigator, again acting as 'student', entered the answers from these two scripts into the direct-answer program and also answered on MCQ sheets which were scanned and checked manually.

Ethical clearance for this study was obtained from the University of KwaZulu-Natal ethics committee (ethical clearance approval number: HSS/0792/08).

RESULTS

The scores of the students' marks as obtained on the paper-based test and the computer program-based test are shown in Table 1.

Table 1:
Scores obtained in paper-based test and computer-based test, shown as actual score out of a total of 150 and as a percentage

Script number	**Paper-based test**				**Result in computer-based test**		**Discrepancy between paper-based and computer-based test**
	Result on scanning MCQ sheet		**Result on manual marking of MCQ sheet**				
	Score /150	**%**	**Score /150**	**%**	**Score /150**	**%**	**%**
001	123	82.0	123	82.0	124	83.0	1.0
002	122	81.3	122	81.3	125	83.3	2.0

Script number	Paper-based test				Result in computer-based test		Discrepancy between paper-based and computer-based test
	Result on scanning MCQ sheet		Result on manual marking of MCQ sheet				
	Score /150	%	Score /150	%	Score /150	%	%
003	122	81.3	122	81.3	125	83.3	2.0
004	90	60.0	90	60.0	94	62.7	2.7
005	86	57.3	86	57.3	89	59.3	2.0
006	78	52.0	78	52.0	83	55.3	3.3
007	56	37.3	56	37.3	62	41.3	4.0
008	56	37.3	56	37.3	65	43.3	6.0
009	42	28.0	42	28.0	53	35.3	7.3
A	0	0	0	0	75	50	50
B	0	0	0	0	19	12.7	12.7

Table 1 indicates that the scores of all the students were higher in the directly-entered computer program-based test. Furthermore, the rank order of scripts 001, 002 and 003, and of 007 and 008 changed. Closer scrutiny of these scripts indicated that, in an item where more than half the options were answered incorrectly, the computer program had rounded the mark of this item to zero, so that the negative mark was not carried over to the final score. For example, where an item contained five options, if three of these options were answered incorrectly, this yielded a score of 0 in the computer-based test whereas manual and scanned marking of the MCQ sheets allocated -1 to that item. Thus manual and scanned marking carried over a negative mark of an item to the final score whereas the marking of the computer-based test did not, resulting in discrepancies in the final scores, which were thus higher in the computer-based test.

In the two experimental scripts, in each of which 50% of the questions were answered correctly and 50% incorrectly, the score obtained from manual marking as well as scanning of the MCQ sheet was 0%. In the direct entry computer-based test, script A, with the correct answers clustered together, however, scored 50%. Script B – with correct and incorrect answers scattered throughout the script – scored 12.7%.

DISCUSSION

This study illustrates how discrepancies in final test scores can occur in computer-based tests which allocate scores by using negative marking strategies. Whilst allocation of negative marks for incorrect answers was one of the options of the new program, not all of these negative marks were reflected in the calculation of final test scores. Investigation of the computer program under consideration revealed that a setting has to be activated prior to running the assessment, in order for each item's negative mark to be added to the final score. If this is not done, a subtotal of zero is assigned to each item with an overall negative score for its components. This is referred to as 'internal negative marking' (Cook, 2010). The result of this manoeuvre is the discrepancies in the scores obtained. We are aware of at least one other computerised assessment program that calculates test scores in the same way.

We are aware too, of other medical schools which deliberately do not carry over negative marks obtained in individual items to the final score, in the belief that marks lost out of ignorance in one area should not adversely affect marks gained from knowledge in another area (TS, personal communication). Negative marking is designed to discourage, or at least to adjust for, random guessing (Holt, 2006). We argue that for the formula to have the desired effect, one mark should be deducted for each incorrect answer. A system that does not carry individual items' negative marks to the final score is less effective in discouraging guessing. If the penalty for guessing is not both sufficient in magnitude and consistently applied, it may allow strategic students to leave out parts of the curriculum when studying. In a given item covering material that they have not studied and for which they may have to guess all the true/false statements, they can count on at least some negative marks not being carried over to the final mark. At the same time, they hope to score well on items that they have studied. This marking scheme conveys a relative advantage on those students with knowledge in only some areas of the syllabus over those students with a broader spread of knowledge in all areas of the syllabus.

Adjusting the pass mark

We are aware of yet another university that has taken a different approach: it uses true/false questions and allocates no negative marks. This has the effect of reducing the useable range of marks from 0-100 to 50-100, thus allowing less discrimination between students of different ability (Pamplett and Farnill, 1995). A mark of 75% then represents a pass. Since our university has decided that the pass mark for all modules is 50%, we are constrained to use negative marking to achieve a mark range of 0-100%.

In general, the higher the pass mark and the greater the number of options, the smaller the impact of random guessing. Thus the former university referred to above has elected to use true/false questions – effectively two options – with a 'random guess score' of 50%. Setting the pass mark at a score of 50% under such circumstances would be ridiculous without negative marking, since a student with no knowledge at all could be expected to answer 50% of the questions correctly by chance. Setting the pass mark at 75% (50% along the useable mark range) is their way of negating the effect of random guessing. Yet another approach would be to avoid true/false questions and to use, say, five-option MCQs; this would allow a score of 20% from random guessing, so setting a pass mark of 50% would be less unreasonable (although a pass mark of 60%, being 50% along the useable mark range, would be more accurate). However, this comes at the cost of a smaller range of meaningful marks, and thus a diminished ability to distinguish between students of differing academic ability.

Inconsistencies arising from 'internal negative marking'

In the computer-based test program under consideration, the problem of individual items' negative marks not being carried over to the final score can be avoided by activation of a setting. The program could thus be useful to us in the future. However, it is of concern that the possibility exists – and may be deliberately chosen – not to carry over each item's negative marks to the final score. As can be seen from our students' scripts, the discrepancy introduced by rounding up items' negative scores to zero is not uniform. It favours those with lower scores – logically, since they have more wrong answers, which attract negative marks. This can be seen as being helpful to weaker students; it can also be seen as giving them – and their examiners – a false sense of their academic ability.

The extreme is illustrated by scripts A and B. Each got 50% of the answers correct, but rounding every question item's negative score to zero is patently unfair, since that results in a discrepancy of 37.3% between the two, with one 'student' passing and the other failing. While script A is an extreme – and probably unrealistic – example, it highlights the same unfairness as illustrated by scripts 007 and 008 in our initial sample. Both of these scored 37.3% with all negative scores included. With rounding up to zero within items, script 007 then scored 41.3% and script 008 scored 43.3%. A 2% difference could represent the difference between one student being granted a supplementary paper and the other – able to answer the same number of questions correctly – not crossing that threshold. Not only is this unfair, it reduces the reliability of the assessment – two students answering 56 questions correctly received different scores. Furthermore scripts 002 and 003 each obtained 81.3% in the original test but this changes to 83.3% in the computerised version. This has now brought these students to the top of the class, as the mark of script 001 changed from 82% to 83%. This further highlights the unfairness of this system.

If one argues that 'students' A and B both deserve to pass because they have answered 50% of the questions correctly, one must not only reconcile their widely disparate scores in the 'item rounding' marking scheme. The fact remains that the number of their correct and incorrect answers cannot be distinguished from those of a person who knows nothing about the assessment material at all and answers at random. From the point of view of their future professions, one has to consider that for the 50% that they 'know', there is another 50% that they *think* they know – but this 'knowledge' is incorrect. It is important for educated people to know their limitations – to know what they know and to know what they do *not* know and with which they need help. The student, who answers only 50% of the questions, but all of them correctly, will arguably make a safer doctor, a more reliable engineer, a more trustworthy accountant, than the one who answers 50% correctly and 50% incorrectly.

CONCLUSION

While the literature correctly identifies several important criteria for an adequate assessment, it can be argued that these requirements relate ultimately to the fairness of the procedure – its ability to reflect accurately the ability of the examinees being assessed. This study does not primarily address the relative validity or reliability of MCQ types compared to other instruments, nor enters into the details of the debate on negative marking. It does affirm that, if negative marking is to be used, particularly for true/false questions, prescribed negative marking should be applied consistently throughout an assessment. From the actual and constructed examples used, we argue that 'item rounding' or 'internal negative marking' is neither reliable nor fair. The theory behind item rounding – not allowing an examinee's lack of knowledge in one area to detract from knowledge in another area – is initially persuasive. However, its outworking in practice detracts from the adequacy of assessment and, for the 'exam-savvy' student, may encourage 'spotting' of certain areas of knowledge and avoidance of others. This study exposes weaknesses not originally apparent, and we argue that the carrying over of *all* negative marks to the final test score is essential for accurate, reliable assessment, fairness to students and fairness to those who rely on their future judgement.

REFERENCES

Akdemir, O. and Oguz, A. (2008) 'Computer-based testing: An alternative for the assessment of Turkish undergraduate students' *Computers & Education* 51 pp.1198-1204.

Bandaranayake, R., Payne, J. and White, S. (1999) 'Using multiple response true-false multiple choice questions' *Australian and New Zealand Journal of Surgery* 69 pp.311-315.

Bauer, D., Holzer, M., Kopp, V. and Fischer, M.R. (2011) 'Pick-N multiple choice-exams: a comparison of scoring algorithms' *Advances in Health Science Education* 16 pp.211-221.

Bernardo, J.M. (1998) 'A decision analysis approach to multiple-choice examinations' In F.J. Giron (Eds.) *Applied Decision Analysis*. Dordrecht, Netherlands: Kluwer Academic.

Betts, L.R., Elder, T.J., Hartley, J. and Trueman, M. (2009) 'Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes?' *Assessment & Evaluation in Higher Education* 34(1) pp.1-15.

Brady, A.M. (2005) 'Assessment of learning with multiple-choice questions' *Nurse Education in Practice* 5 pp.238–242.

Bugbee, J. and Alan, C. (1996) 'The equivalence of paper-and-pencil and computer-based testing' *Journal of Research on Computing in Education* 28(3) pp.282-299.

Burton, R.F. (2004) 'Multiple choice and true/false tests: reliability measures and some implications of negative marking' *Assessment & Evaluation in Higher Education* 29(5) pp.585-595.

Burton, R.F. (2005) 'Multiple-choice and true/false tests: myths and misapprehensions' *Assessment & Evaluation in Higher Education* 30(1) pp.65-72.

Carneson, J., Delpierre, G. and Masters, K. (1996) 'Designing and Managing MCQs' Centre for Educational Technology (CET), University of Cape Town. http://web.uct.ac.za/projects/cbe/mcqman/mcqchp3.html (Accessed 12 May 2012).

Chang, S.H, Lin, P.C. and Lin, Z.C. (2007) 'Measures of partial knowledge and unexpected responses in multiple-choice tests' *Educational Technology & Society* 10(4) pp.95-109.

Cisar, S.M., Cisar, P. and Pinter, R. (2009) 'True/false questions analysis using computerized certainty-based marking tests' *IEEE* pp.171-174.

Collins, J. (2006) 'Education techniques for lifelong learning: Writing multiple-choice questions for continuing medical education activities and self-assessment modules' *Radiographics* 26 pp.543-551.

Cook, J. (2010) 'Getting started with e-assessment' Project Report. Bath: University of Bath. http://opus.bath.ac.uk/17712/1/Getting_started_with_e-assessment_14Jan2010.pdf (Accessed 10 June 2010).

Davies, P. (2000) 'Computerised peer assessment' *Innovations in Education and Training International* 37(4) pp.346-355.

Fowell, S.L. and Jolly, B. (2000) 'Combining marks, scores and grades. Reviewing common practices reveals some bad habits' *Medical Education* 34 pp.785-786.

Goldik, Z. (2008) 'Abandoning negative marking' *European Journal of Anaesthesiology* 25 pp.349-351.

Hammond, E.J., McIndoe, A.K., Sansome, A.J. and Spargo, P.M. (1998) 'Multiple-choice examination: adopting an evidence based approach to exam technique' *Anaesthesia* 53 pp.1105-1108.

Hannan, A, English, S. and Silver, H. (1999) 'Why Innovate: Some preliminary findings from a research project on Innovations in teaching and learning in higher education' *Studies in Higher Education* 24(3) pp.279-289.

Holsgrove, G. (1992) 'Guide to post graduate exams: multiple choice questions' *British Journal of Hospital Medicine* 48(11) pp.757-761.

Holt, A. (2006) 'An analysis of negative marking in multiple-choice assessment' In S. Mann and N. Bridgeman (Eds.) *28th Annual Conference of the National Advisory Committee on Computing Qualifications (NACCQ 2006)*. Wellington, New Zealand.

Imperial College. (2012) *Guidance on using multiple choice questions (MCQ) in assessment.* http://www.studynet2.herts.ac.uk/intranet/lti.nsf/Teaching+Documents/840F6136E87D0594802577A1004820AF/$FILE/Guidance%20on%20Using%20Multiple%20Choice%20Questions%20(MCQ)%20in%20Assessment.pdf (Accessed 23 April 2012).

Jennings, S. and Bush, M. (2006) 'A comparison of conventional and liberal (free-choice) multiple-choice tests' *Practical Assessment, Research & Evaluation* (8) pp.1-5. http://pareonline.net/getvn.asp?v=11&n=8 (Accessed 22 April 2012).

Manogue, M., Kelly, M., Bartakova Masaryk, S., Brown, G., Catalanotto, F., Choo-Soo, T., Delap, E., Godoroja, P., Morio, I., Rotgans, J. and Saag, M. (2002) 'Evolving methods of assessment' *European Journal of Dental Education* 6 (Suppl 3) pp.53–66.

McCoubrie, P. (2004) 'Improving the fairness of multiple-choice questions: a literature review' *Medical Teacher* 26(8) pp.709-712.

McHarg, J., Bradley, P., Chamberlain, S., Ricketts, C., Searle, J. and McLachlan, J.C. (2005) Assessment of progress tests' *Medical Education* 39(2) pp.221-227.

Morrison, S., and Free, K.W. (2001) 'Writing multiple-choice test items that promote and measure critical thinking' *Journal of Nursing Education* 40(1) pp.17-24.

Moss, E. (2001) 'Multiple choice questions: their value as an assessment tool' *Current opinion in anaesthesiology* 14(6) pp.661-666.

Pamplett, R. and Farnill, D. (1995) 'Effect of anxiety on performance in multiple-choice examination' *Medical Education* 29 pp.298-302.

Peak, P. (2005) 'Recent Trends in Comparability Studies' *In Pearson Education Management Research Report 05-05* http://www.pemsolutions.com/downloads/research/TrendsCompStudies_rr0505.pdf (Accessed 10 June 2010).

Smoline, D.V. (2008) 'Some problems of computer-aided testing and "interview-like tests"' *Computers & Education* 51 pp.743-756.

Zakrzewski, S. and Steven, C. (2003) 'Computer-based assessment: quality assurance issues, the hub of the wheel' *Assessment & Evaluation in Higher Education* 28(6) pp.609-623.

Zimmerman, D.W. and Williams, R.H. (2003) 'A new look at the influence of guessing on the reliability of multiple-choice tests' *Applied Psychological Measurement* 27(5) pp.357-371.