Medical Teacher, Vol. 26, No. 8, 2004, pp. 709–712

Taylor & Francis
healthsciences

Improving the fairness of multiple-choice questions: a literature review

PAUL McCOUBRIE
Department of Radiology, Bristol Royal Infirmary, Bristol, UK

Introduction

Multiple-choice questions (MCQs) need little introduction. They are widespread amongst the steady diet of exams fed to medical students and postgraduate trainees. In particular, they are popular for evaluating medical students given the sheer logistical advantage of being able to test large numbers of candidates with minimal human intervention.

It might be claimed that their ease of use and testing efficiency comprise the sole rationale for their continued use. Other substantive criticisms include unfairness (Kaufman, 2001), the promotion of factual regurgitation over higher order thinking and their lack of professional authenticity (van der Vleuten, 1996). Many MCQ formats and other alternatives such as the modified essay question (MEQ) and even the objective structured clinical exam (OSCE) can claim a lineage to dissatisfaction with MCQs (Norman, 1996). Accordingly, a trend away from possible over-reliance on MCQs has been noted in the last two decades, particularly in the undergraduate setting (Fowell *et al*., 2000).

This article will concentrate on developing an evidence-based strategy to use MCQs more fairly so that MCQs can continue to have an important role in assessment and a positive effect on learning.

Review method

The search terms 'assessment', 'examination', 'MCQ' and 'multiple choice question' were applied to Medline (1980 to present), ERIC and TIME. Ancestry searching was performed on the articles thereby identified. The information search was supplemented by information from standard textbooks and the grey literature.

MCQ testing of competence: is it fair?

Possession of an adequate knowledge base was once viewed as unimportant, as knowledge is changing so rapidly. The ability to find out and problem-solve was stressed as being more important. However, problem-solving and competence are not generic and are dependent on and specific to individual cases, tasks, situations, problems and, crucially, are knowledge-dependent (Norman, 1996).

A longstanding criticism of the validity of MCQs is that testing cognitive (or factual) knowledge does not guarantee competence as professional competence integrates knowledge, skills, attitudes and communication skills. However, decades of research into reasoning and thinking have unequivocally shown that knowledge of a domain is the single best determinant of expertize (Glaser, 1984). MCQs are, therefore, a valid method of competence testing, as cognitive knowledge is best assessed using written test forms (Downing, 2002).

The common assumption that MCQs are intrinsically more reliable than other written forms of testing is ill founded. MCQ-based exams are reliable only because they are time-efficient and a short exam still allows breadth of sampling of any topic. Other written examination formats are slower to complete and hence cannot realistically sample as widely unless the test is several hours long.

While MCQs are expressly designed to assess knowledge, well-constructed MCQs can also assess taxonomically higher-order cognitive processing such as interpretation, synthesis and application of knowledge rather than testing recall of isolated facts (Case & Swanson, 2001). However, 'higher-order' MCQs still require cognitive knowledge and may not be any more valid but their realism makes them more acceptable to examinees and examiners (Peitzman *et al*., 1990; Case *et al*., 1996). However, a fair MCQ-based test is much more than a statistically reliable test of cognitive knowledge.

Exam fairness

The number of stakeholders interested in medical assessment has increased. Previously, an exam merely concerned the individual and her/his examiners. Now assessment plays an increasing role in satisfying quality issues of registration bodies and employers, as well as reassuring the public. This is particularly true in 'high-stakes' exams, where life-altering

Correspondence: Paul McCoubrie, Specialist Registrar in Radiology, Department of Radiology, Bristol Royal Infirmary, Marlborough Street, Bristol BS2 8HW, UK. Email: pmccoubrie@doctors.org.uk

ISSN 0142–159X print/ISSN 1466–187X online/04/000709-4 © 2004 Taylor & Francis Ltd
DOI: 10.1080/01421590400013495

decisions are made concerning graduation and career progression.

The fairness of an exam is a judgement of its authenticity by an exam 'stakeholder', something more substantial than face validity. This global impression reflects the faith established by not only adequate psychometric qualities (mainly reliability and validity) but also due diligence of construction together with appropriate pass/fail standard setting. This integrated concept has recently attracted more attention (Kaufman, 2001) and equates to the features of a legally defensible exam (Downing, 2002).

Senior medical students and postgraduate trainees become highly experienced in taking exams. Trainees' perceptions of fairness of an exam are valid and should be sought during evaluative processes (Duffield & Spencer, 2002). It is unfortunately difficult to judge the fairness of many medical exams, as UK universities and Royal Colleges have been historically reluctant to publish any exam data (Hutchinson *et al*., 2002).

Consequential validity

Assessment drives learning in at least four ways: its content, its format, its timing and any subsequent feedback given to the examinee (van der Vleuten, 1996). The 'law' of educational cause and effect states that:

> For every evaluative action, there is an equal (or greater) (and sometimes opposite) educational reaction. (Schuwirth, 2001)

Phrased metaphorically, 'the assessment tail wags the curriculum dog' or more crudely, ''grab the students by the tests and their hearts and minds will follow'' (Swanson & Case, 1997). The real art comes in using assessment methods to steer students' learning appropriately. This aspect of an examination is referred to as its *consequential validity*. For example, an examinee will prepare more thoroughly if an exam is high-stakes and perceived to be difficult. However, examinees have a finite amount of time, tests are generally tough and learning is therefore strategic (Coles, 1998). Motivation is complex but socially based value judgements on the status of a subject should not be overlooked (Sinclair, 1997).

If an exam is fair then students are more prone to study the subject rather than studying the exam and vice versa. The phenomenon of studying the exam rather than the subject is well known amongst examinees. Several 'cue-seeking' techniques such as topic spotting and exhaustive MCQ practice have been described (Williams *et al*., 1996).

Consequential validity is an increasingly important concept but is often ignored by examiners. A recent review of postgraduate certification processes found that only two of 55 papers addressed the education impact of assessments on learning (Hutchinson *et al*., 2002). The 'Hidden Curriculum' is a useful model to understand the crucial effect that assessment has on the interaction between the curriculum, teaching and learning (Synder, 1971). In this, the examination programme becomes the hidden or 'real' curriculum. This can be minimized by aligning assessment to curricular goals and teaching/learning activities.

MCQs and, indeed, all written assessments lack professional authenticity and are arguably less valid as a result. Competence is contextual and recall to written test items may be impaired (van der Vleuten, 1996). Over-reliance on written forms of assessment can lead to unforeseen, unwanted educational consequences such as over-reliance on written learning (Newble & Jaeger, 1983). So, to make testing both fair and consequentially valid, MCQs should be used strategically to test important content, and mixed with practical testing of clinical competence.

Assessment by ambush

One aspect of an unfair exam is the so-called 'assessment by ambush' (Brown, 1992). Here the choice of questions is determined by the desire to trip the unwary and discriminate as clearly as possible between high and low achievers. This quest for discrimination can lead to the deliberate omission of questions on essential parts of the curriculum because they are 'too easy' and insufficiently discriminatory. Examinees are consequently driven to learn minutiae, skipping over potentially important topics.

There are two issues at stake here:

(1) *Improper sampling*: To sample content adequately, testing needs to cover a broad range of topics. Furthermore, some effort should be made to address important content. This may all be done with rigid assessment design techniques such as 'blueprinting' where content is sampled deliberately (Jolly, 1999).
(2) *Cueing effect*: True/false format MCQs inadvertently provide cues, resulting in less discriminatory questions (Veloski *et al*., 1999).

Question construction

The fairness concept also encapsulates technical aspects of question construction. Focusing on important content is not sufficient. Crucial content cannot be assessed unless the question is simple and easy to understand by being well structured and free of construction errors (Case & Swanson, 2001). Subtle grammatical chicanery, particularly the use of negatives and imprecise terms (e.g. 'frequently') may cause confusion amongst examinees (Case, 1994; Holsgrove & Elzubeir, 1998). Any confusion over grammar or question structure invalidates the test as (1) this extra grammatical variable does not relate to knowledge of the subject, (2) it discriminates against examinees for whom English is not their first language and (3) benefits the wily and experienced examinee.

Standard setting

Pass/fail standard setting is a judgemental process that will always be based on an arbitrary decision related to the point on the continuum of competence where one separates the competent from the incompetent (Cusimano, 1996). Furthermore, ''the process of setting performance standards is open to constant criticism and remains controversial to discuss, difficult to execute and almost impossible to defend'' (Berk, 1986).

A criterion-referenced standard (analogous to a driving test) is preferred to a norm-referenced (fixed pass rate) or holistic model (arbitrary pass mark at, say, 60%) (Case & Swanson, 2001). Norm-referenced or holistic models are in common usage, yet are the least defensible in high-stakes

exams of clinical competence. At least one UK Royal College has adopted criterion-referenced standards (Lowe *et al*., 1998). However, over 35 criterion-referenced methods of standard setting have been proposed (Berk, 1986). Radically different pass marks can be obtained depending on the method used, varying by a factor of 42 (Cusimano, 1996).

There is no single recommended approach to setting standards (Ben-David, 2000). For MCQ papers, Case & Swanson (2001) advise the use of content-based procedures (e.g. the modified Angoff procedure) or a compromise model (e.g. Hofstee). Content-based procedures are relatively simple but not in widespread use as compromise models are less time-consuming and probably equally reliable (D.I. Newble, personal communication, 2003).

Reusing MCQ questions

Computer-marked MCQs calculate statistical markers of each question's difficulty and discriminatory capacity. Discriminatory questions of moderate difficulty are often retained and reused. However, the question may be difficult or discriminate because it contains inappropriate content or grammatical flaws. Therefore, each one should be critically reviewed before reuse. Even carefully designed items benefit from rigorous evaluation and 30% to 60% warrant revision or deletion (Fajardo & Chan, 1993; Dixon, 1994). Also, scores on single questions on a single occasion are poor predictors of overall performance, and weighting of scores on such discriminatory and difficult questions is unfair.

Most institutions maintain a confidential bank of MCQ questions that are reused to a varying degree. There is slight controversy over reusing MCQs. One study claimed students do not remember or pass on these questions (Herskovic, 1999). However, the experience of high-stakes examiners (Lowe *et al*., 1998) is that examinees transmit old questions, necessitating constant renewal of the MCQ bank. Excessive reuse of MCQs may be unfair as not all examinees will have had exposure to old questions. Copying and circulation of MCQs by examinees could be construed as unacceptable, perhaps even cheating, but is probably unavoidable. Some responsibility for this must lie with the medical academic establishment, with ''overemphasis on grades and competition [and] … a hidden curriculum which delivers negative messages'' (Glick, 2001).

Fairness and MCQ formats

The basic MCQ model comprises a lead-in question (stem) followed by a number of answers (branches). There are two major formats.

True/False format. In the UK, these are known as 'multiple true/false' questions and are the most common format (Fowell *et al*., 2000). This model suffers from two major drawbacks, guessing and the cueing effect. Discouragement of guessing is often necessary and usually achieved by negative marking. However, guessing ability is unrelated to the subject being tested (Jolly, 1999). Negative marking introduces this additional variable, leading to negative psychometric effects (Schuwirth *et al*., 1996a). Cueing can be difficult to disentangle from guessing but has been estimated to play a role in approximately 20% of answers (Schuwirth *et al*., 1996b).

Case & Swanson (2001) cite different reasons as to why this format has been abandoned in the USA. After reviewing 'literally tens of thousands' of true/false MCQs, they found that they are not only difficult to write well but, in order to avoid ambiguity, the writer is pushed to assessing the recall of an isolated fact. Such a format is therefore unfair as an otherwise competent student may fail if he/she has not memorized isolated facts.

Single-best answer family. This format includes a variety of formats where the one or more correct response(s) is/are selected from a list of possibilities. Focus has shifted from traditional 3–5 branches to larger numbers of branches. This may be 20–30 in the case of extended-matching questions (EMQs), or up to 500 for open-ended or 'uncued' formats (Veloski *et al*., 1999). Larger numbers of branches effectively eliminate cueing but there seems little advantage in increasing this number over 20, as there is no further reduction in the cueing effect (Fajardo & Chan, 1993; Fenderson *et al*., 1997). EMQs are more difficult, more reliable, more discriminating, and allow testing time to be reduced in addition to being quicker and easier to write than other formats (Case & Swanson, 1993; Beullens *et al*., 2002).

Computer-based MCQs

Computer-based MCQ testing has been described for nearly 20 years and many US licensing exams are now to some extent computerized. It can have advantages in terms of costs, development and delivery, and it enables the use of electronic multimedia (Clauser & Schuwirth, 2002). Fears about 'computer anxiety' affecting performance seem to be ill founded (Schuwirth *et al*., 1996a; Lee & Weerakoon, 2001).

Like a good oral exam or, indeed, a high-jump competition, Computerized Adaptive Testing (CAT) targets the test to the individual, and consequently can be seen as fairer. Although it is costly to set up and maintain a computerized bank of validated questions, CAT is being increasingly adopted in large-scale testing programs, as testing times may be reduced dramatically (Wise & Kingsbury, 2000). The Medical Council of Canada recently switched to an Internet-based CAT licensing examination and reduced testing time from two days to 3½ hours, thereby reducing examinee fatigue whilst maintaining acceptable levels of reliability (Miller *et al*., 2002).

Notes on contributor

PAUL MCCOUBRIE BSc (Hons) MBBS MRCP FRCR, is a year five Specialist Registrar in Radiology who is also studying for an MEd at the University of Sheffield. His conflict of interest is personal suffering for over 13 years at the hands of poorly constructed MCQs in medical examinations.

References

BEN-DAVID, M.F. (2000) AMEE Guide No. 18: Standard setting in student assessment, *Medical Teacher*, 22, pp. 120–130.

BERK, R.A. (1986) A consumer's guide to setting performance standards on criterion-referenced tests, *Review of Educational Research*, 56, pp. 137–172.

BEULLENS, J., VAN DAMME, B., JASPAERT, H. & JANSSEN, P.J. (2002) Are extended-matching multiple-choice items appropriate for a final test in medical education, *Medical Teacher*, 24, pp. 390–395.

BROWN, S. (1992) Trends in assessment, in: R. Harden, I. Hart & H. Mulholland (Eds) *Approaches to the Assessment of Clinical Competence*, Vol. 1 (Dundee, Centre for Medical Education), pp. 3–8.

CASE, S.M. (1994) The use of imprecise terms in examination questions: how frequently is frequently?, *Academic Medicine*, 69, pp. S4–S6.

CASE, S.M. & SWANSON, D.B. (1993) Extended matching items: a practical alternative to free response questions, *Teaching and Learning in Medicine*, 5, pp. 107–115.

CASE, S.M. & SWANSON, D.B. (2001) *Constructing Written Test Questions for the Basic and Clinical Sciences*, 3rd edn (Philadelphia, National Board of Medical Examiners).

CASE, S.M., SWANSON, D.B. & BECKER, D.F. (1996) Verbosity, window dressing and red herrings: do they make a better test item?, *Academic Medicine*, 71, pp. S28–S30.

CLAUSER, B.E. & SCHUWIRTH, L.W.T. (2002) The use of computers in assessment, in: G. Norman, C. van der Vleuten & D. Newble (Eds) *International Handbook of Research in Medical Education*, Vol. 2 (Dordrecht, Kluwer), pp. 757–792.

COLES, C.R. (1998) How students learn: the process of learning, in: B. Jolly & L. Rees (Eds) *Medical Education in the Millennium* (Oxford, Oxford University Press), pp. 63–82.

CUSIMANO, M.D. (1996) Standard setting in medical education, *Academic Medicine*, 71, pp. S112–119.

DIXON, R.A. (1994) Evaluating and improving multiple choice papers: true–false questions in public health medicine, *Medical Education*, 28, pp. 400–408.

DOWNING, S.M. (2002) Assessment of knowledge with written test formats, in: G. NORMAN, C. VAN DER VLEUTEN & D. NEWBLE (Eds) *International Handbook of Research in Medical Education*, Vol. 2 (Dordrecht, Kluwer), pp. 647–672.

DUFFIELD, K.E. & SPENCER, J.A (2002) A survey of medical students' views about the purposes and fairness of assessment. *Medical Education*, 36, 879-886.

FAJARDO, L.L. & CHAN, K.M. (1993) Evaluation of medical students in radiology: written testing using uncued multiple choice questions, *Investigative Radiology*, 28, pp. 964–968.

FENDERSON, B.A., DAMJANOV, I., ROBESON, M.R., VELOSKI, J.J. & RUBIN, E. (1997) The virtues of extended matching and uncued tests as alternatives to multiple choice questions, *Human Pathology*, 28, pp. 526–532.

FOWELL, S.L., MAUDSLEY, G., MAGUIRE, P., LEINSTER, S.J. & BLIGH, J. (2000) Student assessment in undergraduate medical education in the United Kingdom 1998: report of findings, *Medical Education*, 34, pp. 1–61.

GLASER, R. (1984) Education and thinking: the role of knowledge, *American Psychology*, 39, pp. 193–202.

GLICK, S.M. (2001) Cheating at medical school, *British Medical Journal*, 322, pp. 250–251.

HERSKOVIC, P (1999) Reutilization of multiple-choice questions. *Medical Teacher*, 21, 430-431.

HOLSGROVE, G. & ELZUBEIR, M. (1998) Imprecise terms in UK medical multiple-choice questions: what examiners think they mean, *Medical Education*, 32, pp. 342–350.

HUTCHINSON, L., AITKEN, P. & HAYES, T. (2002) Are medical postgraduate certification processes valid? A systematic review of the published evidence, *Medical Education*, 36, pp. 73–91.

JOLLY, B. (1999) Assessment and examination, *Advances in Psychiatric Treatment*, 5, pp. 405–414.

KAUFMAN, D.M. (2001) Assessing medical students: hit or miss, *Student BMJ*, 9, pp. 87–88.

LEE, G. & WEERAKOON, P. (2001) The role of computer-aided assessment in health professional education: a comparison of student performance in computer-based and paper-and-pen multiple choice tests, *Medical Teacher*, 23, pp. 152–157.

LOWE, D.G., FOULKES, J. & RUSSELL, R.C.G. (1998) ABS to MRCS at the RCS: philosophy, format and future, *Annals of the Royal College of Surgeons of England*, 80, pp. 213–218.

MILLER, D.R., BLACKMORE, D.E., BOULAIS, A.-P. & DAUPHINEE, W.D. (2002) The challenges of creating a computer program designed to administer a high-stake licensing examination, in: *AMEE 2002: Approaches to Better Teaching (Conference Proceedings)* Lisbon, Portugal, pp. 4–105.

NEWBLE, D.I. & JAEGER, K. (1983) The effect of assessment and examinations on the learning of medical students, *Medical Education*, 13, pp. 263–268.

NORMAN, G.R. (1996) Multiple choice questions, in: S. Shannon & G. Norman (Eds) *Evaluation Methods: A Resource Handbook*, 3rd edn (Hamilton, Canada), pp. 47–54.

PEITZMAN, S.J., NIEMAN, L.Z. & GRACLEY, E.J. (1990) Comparison of fact recall with high order questions in multiple choice examinations as predictors of clinical performance of medical students, *Academic Medicine*, 65, pp. S59–S60.

SCHUWIRTH, L.W.T. (2001) General concerns about assessment [available online at: www.fdg.unimaas.nl/educ/lambert/ubc, accessed 12 December 2002].

SCHUWIRTH, L.W.T., VAN DER VLEUTEN, C., STOFFERS, H.E.J.H. & PEPERKAMP, A.G.W. (1996a) Computerized long-menu questions as an alternative to open-ended questions in computerized assessment, *Medical Education*, 30, pp. 50–55.

SCHUWIRTH, L.W.T., VAN DER VLEUTEN, C.P.M. & DONKERS, H.H.L.M. (1996b) A closer look a cueing effects in multiple-choice questions, *Medical Education*, 30, pp. 44–49.

SINCLAIR, S. (1997) *Making Doctors: An Institutional Apprenticeship* (Oxford, Berg).

SWANSON, D.B. & CASE, S.M. (1997) Assessment in basic science instruction: directions for practice and research, *Advances in Health Sciences Education*, 2, pp. 71–84.

SYNDER, B.R. (1971) *The Hidden Curriculum* (New York, Knopf).

VAN DER VLEUTEN, C. (1996) The assessment of professional competence: developments, research and practical implications, *Advances in Health Sciences Education*, 1, pp. 41–67.

VELOSKI, J.J., RABINOWITZ, H.K., ROBESON, M.R. & YOUNG, P.R. (1999) Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence, *Academic Medicine*, 74, pp. 539–546.

WILLIAMS, C., TRIGWELL, P. & YEOMANS, D. (1996) MCQ technique, *British Journal of Hospital Medicine*, 55, pp. 479–481.

WISE, S.L. & KINGSBURY, G.G. (2000) Practical issues in developing and maintaining a computerized adaptive testing program, *Psicologica*, 21, pp. 135–155.