

## MEDICAL TEACHER 25TH ANNIVERSARY SERIES

# Multiple-choice questions revisited

JOHN ANDERSON

*University of Newcastle upon Tyne, UK*

**SUMMARY** MCQs of the multiple true/false (MTF) variety were widely used in summative assessment 25 years ago. They could test a number of skills in addition to recall of factual knowledge, and were reliable, discriminatory, reproducible and cost-effective. However, there are now considerable doubts about their construct validity, mainly because of the varying responses of examinees to negative countermarking and the 'don't know' option, and the strategies they use when sitting examinations. Extended matching and one-from-five questions are now preferable, and negative countermarking is outmoded. MTF questions are still valuable in formative assessment and revision but are not recommended for summative examinations.

## Introduction

Twenty-five years ago *Medical Teacher* published two articles in its *Controversy* series; one by me, 'For multiple choice questions' (Anderson, 1979), and one by Sir George Pickering, 'Against multiple choice questions' (Pickering, 1979). Some time later I reviewed the MCQ controversy in the same journal (Anderson, 1981). Sir George's contribution was foreshadowed by the earlier publication of his admirable book, *Quest for Excellence in Medical Education* (Pickering, 1978). Our two articles produced quite a lively correspondence but I think it fair to say that no firm conclusions were drawn one way or the other.

It is a pleasure to be invited to contribute this paper on the Silver Jubilee of *Medical Teacher*. Sadly, Sir George is no longer with us but I think he might have agreed that the way MCQs have evolved in the UK since their first use almost 50 years ago is an example of his pursuit of excellence in medical education.

I was invited to choose my own title for this paper; that which I have used allows me to be suitably reflective and discursive, as well as to place on record my current views on MCQs. Call it my apostasy, if you like!

In this paper I will review the position 25 years ago and then consider how this has evolved into the present state before briefly concluding.

## Then...

### *The background: question types*

Twenty-five years ago the multiple (or independent) true/false (MTF) format was that favoured in the great majority of undergraduate and postgraduate examinations in the UK. The one-from-five (one best response) type was occasionally

used but these questions were then seldom of good quality. Attempts seem seldom to have been made to find items that were plausible 'negative distractors', the key feature of genuine one-from-five questions.

There were, even then, many other MCQ types available. Some of these, particularly the relationship-analysis type, had merit and were worthy of wider use, but in the UK a quarter of a century ago none challenged the primacy of the MTF format. My original paper, and my subsequent Review article, was based on these questions. I emphasized what I regarded as their strengths and discussed their undoubted weaknesses, but I did not challenge some of the issues now regarded as contentious, such as the negative scoring of incorrect answers.

### *Strengths of MTF questions*

Although I agreed that these basically test recall of factual knowledge, I argued that they could also test higher taxonomic skills, such as understanding, reasoning, data interpretation and problem-solving (Charvat *et al.*, 1968). I admitted that they could not test the ability of students to create a new synthesis, and my comments dealt only with the cognitive domain; MCQs could not test in the psychomotor and affective areas.

It was generally accepted 25 years ago that MTF questions, when carefully prepared, were:

- reliable;
- discriminatory;
- reproducible;
- cost-effective.

They had good concurrent and predictive validity, and in most respects they ranked highly on the index of utility described later by Van der Vleuten (Van der Vleuten, 1996).

### *Weaknesses of MTF questions*

A number of weaknesses were evident even then. I discussed these at some length. The *face validity* of MTF questions was only satisfactory if the questions were 'good' ones (although the frequent complaints by students of 'ambiguous wording' were often an excuse for lack of knowledge or a reluctance to use reasoning!). As to *content validity*, there was sometimes a tendency to concentrate on trivia. We also know now that

Correspondence: email: andersgos@jander.demon.co.uk

there were major problems with *construct validity*, mainly related to the marking scheme employed—more of this later.

According to some critics, examinees in MCQ tests were required only to recognize the correct answers or to eliminate the incorrect ones—the *cuing effect*; the active generation of responses was not required. Later research, however, has shown that the effect of cuing is marginal (Schuwirth *et al.*, 1992).

A further criticism was that MTF questions led, almost inevitably, to a *norm-referenced* system of assessment, in that their principal strength was to rank examinees accurately and fairly. The superior criterion-referenced system was difficult to achieve, not least because of problems in agreeing the criteria to be used.

*Standard setting*, particularly setting the pass mark and other cut-off points, was an arbitrary process, often resolved by simply agreeing the percentage of candidates who should pass.

Finally, there were widespread concerns about the effects of MCQs on *learning styles*; the type of question that required fairly low-level recall of factual knowledge seemed likely to encourage surface learning, with students trying to commit isolated facts to memory rather than to understand the topic in depth.

#### *Strengthening the weaknesses*

Recognizing these problems, examiners even then made strenuous efforts to overcome them. These included:

- (1) the *setting and review of questions* by a multispecialty panel of experts to ensure face and content validity;
- (2) the generous and regular *use of 'marker' questions* which had been previously used and validated, thus allowing comparisons of successive cohorts of examinees and an assessment of the overall degree of difficulty of papers, allowing for possible adjustment of standards and cut-off points accordingly;
- (3) the use of *short-scale mark conversion, harmonization of marks and standard scores*—but since these were all dependent on a normal distribution of scores and standard deviations (thus emphasizing the norm-referenced nature of the tests) these methods were probably more cosmetic than curative;
- (4) the definition of an *acceptable level of performance (ALP)*, in an attempt to resolve the problem of setting the pass mark—but this approach was also essentially arbitrary and could not mask the basic problem of norm-referencing;
- (5) as to *patterns of learning*, the fact is that students learn what they are tested on (or think they will be tested on) and are likely to ignore what they are not. This is the 'hidden curriculum'; examinations define academic success and who can blame students for trying to optimize their chances of achieving success? The challenge for those who develop tests has always been to use this phenomenon strategically and to reinforce desirable patterns of learning. It is possible to write higher-level MCQs that demand analysis and problem solving, so the argument about surface learning must be seen in the context of the type of MCQ used;
- (6) in general, it was agreed that MCQs should not be used as a *sole* assessment method in summative examinations,

but should be *used alongside other test forms*. This was designed to broaden the range of skills that would be tested. The use of a battery of tests assessing achievement in all three domains (cognitive, psychomotor and affective) and at all taxonomic levels would also encourage deep learning. There was, though, one exception to this general rule: MCQ papers could be used as a screen or filter, successful candidates being allowed to proceed to a more comprehensive assessment. Screening on the basis of factual knowledge may seem crude, but it works. Knowledge is basic; without it higher taxonomic skills cannot be developed or demonstrated.

#### *Scoring schemes for MCQs*

I hinted earlier that some of the marking schemes used for MTF questions might be responsible for concerns about their construct validity. The response sheets most commonly used 25 years ago allowed examinees three possible choices for each item: 'true', 'false' and 'don't know'. The 'don't know' option allowed for an honest answer and was designed to discourage guessing—not considered to be a wise strategy in medical practice.

The scoring schemes adopted were based on these three options—that most commonly used awarded one mark for a 'correct' answer (whether true or false), deducted one mark for an 'incorrect' answer, and zero for 'don't know'. With this +1, -1, 0 scoring scheme guessing not only was discouraged but candidates who guessed wrongly were penalized.

But whilst guessing may not be a good strategy in medicine, reasoning and weighing up probabilities to reach the correct answer are to be commended. Unfortunately, whilst the use of the 'don't know' option and negative countermarking had the desired effect of discouraging guessing, they also had the unwanted effect of inhibiting reasoning, leading to successive generations of cautious candidates who did not wish to risk losing marks. There is ample evidence that examinees' personalities have a significant effect on their performance and that this is related to the presence of the 'don't know' option and the practice of negative countermarking. In brief, bold and 'testwise' candidates benefit; cautious candidates do not (Sanderson, 1973; Harden *et al.*, 1976; Jolly, 1976; Fleming, 1988).

I frequently lectured and wrote on how to approach MCQ examinations (Anderson, 1982a, 1982b)—effectively describing strategies designed to maximize scores. But since strategies and personalities—as well as the form of the response sheets and the scoring systems used—could clearly affect outcomes, serious doubts were thrown on the construct validity of MTF MCQs. The final scores in an examination were a compound of knowledge and confidence, and it was therefore difficult to reach a reliable conclusion about either.

#### *Negative marking or not?*

The computer can, of course, easily be programmed to award one mark for a correct answer and zero for one that is not correct, whether 'wrong' or 'don't know'. The Royal College of Obstetricians and Gynaecologists ceased negative countermarking of their MTF MCQ papers some years ago, with considerable success. The judicious use of validated 'marker'

questions coupled with the fact that the final scores indicated what candidates actually knew (or were prepared to admit to knowing) formed the basis for a criterion-referenced system. But getting the right answer by chance or guesswork was rewarded; getting it wrong was not penalized—a ‘hidden bonus’. Candidates could gain marks but never lose them, favouring the bold rather than the cautious examinee. Doubts, once more, about construct validity.

### ... and now: the present position

I stand by most of the claims I made 25 years ago regarding the strengths and utility of MCQs and the range of cognitive skills they can reliably test. I accept, though, that there are now much better ways to test data interpretation and problem-solving, for example, than MTF MCQs. Furthermore, my concerns about the limitations of this MCQ type have been considerably strengthened over the years.

In recent years, however, MTF questions have been used less and less in professional examinations, particularly in the postgraduate field, and other objective assessments employed. All are familiar with the modified essay question (MEQ), for example, and in the MCQ field alternative formats have either been resurrected or developed. The Part I MRCP (UK) examination, previously a bastion of MTF questions, now uses the one-from-five type, as does the Membership examination of the Royal College of General Practitioners. Extended matching questions (or items; EMQs or EMIs) are used increasingly in both undergraduate and postgraduate assessment, notably the MRCP and the Part I PLAB examinations. EM questions are invariably set and reviewed by multispecialty panels of experts, and in the Part I PLAB examination (and others) the content of the papers is based on an agreed ‘blueprint’ to ensure content validity. Good one-from-five and EMQs are not easy to set and the latter are complex in form but these questions undoubtedly test much more than recall of factual knowledge. Reasoning, deduction and the intelligent application of well-understood principles and probabilities pay much bigger dividends than plain memory work. Such questions have a high utility index and show better construct validity than MTF MCQs.

Negative countermarking is now seldom used in postgraduate examinations and I welcome this, despite lingering concerns about the possibility of a ‘hidden bonus’. This is, however, much less likely to occur with one-from-five and EMQs than with MTF items; in the latter case the chances of guessing the correct answer are evens, whereas one-from-five questions have a chance of 1 in 5. I would reckon that the chances of guessing the correct answers in the case of EMQs are so low as to be insignificant, because of the structure of the questions.

MCQ papers of the ‘modern’ type are criterion-referenced, and the Angoff method (Angoff, 1971) is regularly used to set standards, thus eliminating or at least minimizing some of the problems that concerned us a quarter of a century ago.

### The current place of MCQ examinations

I remain of the view that, whatever the question type used, MCQ papers perform best as one of a battery of tests in

summative assessments. I still feel, however, that examinations comprising only MCQ papers can be valid if the test is being used as a screen or filter, for the reasons I gave earlier.

As to the type of questions used, the MTF format may still have a place, provided that the +1, 0 scoring system is used. However, I would freely admit that one-from-five questions and EMQs of the type used nowadays are superior.

Where, then, does this leave the MTF question type? The concerns are not so much about the format of the questions, their utility or the range of skills they test as about how they should be scored and the results interpreted. In other words, their use in summative assessment. This clearly gives us a clue. These questions are best suited for situations where scoring, marking and referencing are not required: self-assessment, revision and the identification of learning needs—all aspects of *formative* assessment. This is relevant during all phases of medical education, whether undergraduate, postgraduate or continuing. These are areas where MTF questions, carefully prepared and with due regard to face and content validity, are still of considerable value. But their continued use in summative assessment is not recommended.

### Conclusions

In view of what is known and practised now, were those of us who championed MCQs, particularly of the MTF format, so enthusiastically and, apparently, so uncritically a quarter of a century ago wrong or, at best, misguided? Are we now hopelessly out of date and was the work we did redundant? I don’t think so: education, and medical education in particular, is an evolving and progressive process as new methodologies are tried, tested and developed or discarded. Those of us who were active in medical education a quarter of a century ago built on the work of those who preceded us. If those who follow us do not further refine and develop what we achieved—sometimes proving us wrong in the process—they are unworthy of the trust we have placed in them.


### Notes on contributor

JOHN ANDERSON specialized in endocrinology and diabetes mellitus after qualifying in Newcastle. In 1968 he was appointed Senior Lecturer in Medicine in the University of Newcastle upon Tyne. In 1985 he was appointed as full-time Postgraduate Dean and Professor of Medical Education. He retired in 1998 and his title is now Emeritus. He is a Vice-President of ASME, having previously been Chairman of Council, Secretary and Treasurer.


### References

- ANDERSON, J. (1979) For multiple choice questions, *Medical Teacher*, 1, p. 37.
- ANDERSON, J. (1981) The MCQ controversy—a review, *Medical Teacher*, 3, p. 150.
- ANDERSON, J. (1982a) *The Multiple Choice Question in Medicine*, 2nd edn (London, Churchill Livingstone).
- ANDERSON, J. (1982b) How to tackle multiple choice question papers, *Hospital Update*, 8, p. 593.
- ANGOFF, W.H. (1971) Scales, norms and equivalent scores, in: R.L. Thorndike (Ed.) *Educational Measurement*, 2nd edn (Washington DC, American Council on Education).


- CHARVAT, J., MCGUIRE, C. & PARSONS, V. (1968) *A Review of the Nature and Uses of Examinations in Medical Education* (Geneva, World Health Organization).
- FLEMING, P.R. (1988) The profitability of 'guessing' in multiple choice question papers, *Medical Education*, 22, p. 509.
- HARDEN, R.M., BROWN, R.A., BIRAN, L.A., DALLAS ROSS, W.P. & WAKEFORD, R.E. (1976) Multiple choice questions: to guess or not to guess, *Medical Education*, 10, p. 27.
- JOLLY, B. (1976) Guessing in multiple choice questions, *Medical Education*, 10, p. 530.
- PICKERING, SIR GEORGE (1978) *Quest for Excellence in Medical Education* (London and Oxford, OUP for the Nuffield Provincial Hospitals Trust).
- PICKERING, SIR GEORGE (1979) Against multiple choice questions, *Medical Teacher*, 1, p. 84.
- SANDERSON, P.H. (1973) The 'don't know' option in MCQ examinations, *British Journal of Medical Education*, 7, p. 25.
- SCHUWIRTH, L.W.T., VAN DER VLEUTEN, C.P.M. & DONKERS, H.H.L.M. (1992) Open ended questions versus multiple choice questions: an analysis of cuing effects, in: R.M. Harden, I. Hart & H. Mulholland (Eds) *Approaches to Assessment of Clinical Competence—Part II*, pp. 486–491 (Norwich, Page Brothers).
- VAN DER VLEUTEN, C.P.M. (1996) The assessment of professional competence: developments, research and practical implications, *Advances in Health Sciences Education*, 1, p. 41.



12-14 May 2004  
Sestri Levante,  
Portofino Coast  
Italy



eLearning Results 2004  
— Research & Standards Linked Learning Technologies Summit —



**AMEE**

## eLearning Results 2004

**12-14 May**  
**Sestri Levante, Italy**

The Association for Medical Education in Europe (AMEE) is pleased to be the Medical Education Sponsor of this Meeting, organised in collaboration with Giunti Interactive Labs and IMS Global Learning Consortium, to discuss standards in e-learning.

The programme consists of plenary and large group sessions. The Medical Education session, organised by AMEE, will address the relevance and application of IMS and SCORM standards to medical education and will consist of a series of short case studies followed by interactive discussions.

Open conference sessions take place on 13-14 May and registration is free. Pre-conference workshops will be held on 12 May (charge will be applied for workshops). Places for both open sessions and workshops are strictly limited – pre-registration essential.

**Further information:** <http://www.amee.org/eLearningResults/index.html>  
[p.m.lilley@dundee.ac.uk](mailto:p.m.lilley@dundee.ac.uk)

**To register:** <http://www.elearningresults.com/>