The multiple choice examination paper

G. A. SMART

BRITISH POSTGRADUATE MEDICAL FEDERATION. LONDON

It has been repeatedly shown that under the conditions in which conventional written examinations of the essay type are usually marked there is little internal consistency, that is there may be little or no correlation between the marks obtained by candidates on, say, one half of the paper and those obtained on the other half. Closer scrutiny has also frequently shown a poor correlation between the marks allocated by one examiner and those independently allocated by another (Hartog and Rhodes, 1936). Indeed, a very poor correlation has even been found for the marks awarded to the same set of answers by the same examiner after an interval of several months (Bull, 1956). A very useful review has been published by Cox (1967).

A great deal of this inconsistency lies in the multiplicity of attributes that are tested in a typical essay question, in the very great difficulty of disentangling these attributes, and in the different weighting given to them by different examiners. Moreover, although setting an essay paper is a relatively simple procedure, correcting one is not only highly complex and very subjective, but it is also so time consuming that it is almost impossible for even one particular examiner to maintain the same standard for the duration of his necessarily prolonged marking period. Almost the whole of an examiner's effort goes into the process of awarding marks, and (as is inherent in the method) this activity is carried out with manifest inefficiency.

By contrast, in an efficient multiple choice test nearly all the skill of the examiner is devoted to producing the test material. The scoring of this material is simplified so that it can if necessary be carried out by machine. There can be no variation in the scoring of such a paper and tests of the internal consistency show this to be satisfactorily high.

The range of knowledge that can be tested in a 3-hour multiple choice paper is quite large and certainly reduces the element of luck inherent in an essay paper containing only four to six questions. If such an objective-type paper is set in an effective way, any test material included has to be accepted by the consensus of the examiners as a whole as being both appropriate and technically correct. Irreconcilable differences between examiners on these matters result in the test material being discarded.

Since the marks scored are not influenced by the foibles of any particular person, objective-type questions can be used to play an important part in the process of acquiring knowledge. In the field of cognition good comparisons can be made between different groups so that the effect of different teaching methods can be assessed. Perhaps more important, the technique can be used to design a diagnostic instrument to be used personally by the student. A personal tutorial system of the "Oxbridge" type is almost certainly better, but alas in most medical schools and hospitals it is impracticable.

The virtues of objective testing must, however, not be allowed to blind one to its limitations. The fact is that each attribute that is considered important needs to be tested by methods specifically designed for the purpose. Properly designed multiple choice papers are certainly good for testing factual knowledge and they can test some aspects of the candidate's understanding and use of such knowledge. They cannot test clinical skills nor a whole host of the other desirable assets that should be possessed by a doctor. However, it is quite certain that a doctor who has only a rudimentary knowledge of his subject cannot be really effective, however good his bedside manner.

Types of objective test

In the field of medicine the development of objective tests has been pioneered by the National Board of Medical Examiners in the USA and initially they adopted a number of different formats (Hubbard and Clemens, 1961). Some of these were quite complicated and required skills at unravelling test procedures, which were not relevant to the functions of a doctor. Consequently, the National Board now restricts the tests it uses to three basic types (Hubbard, 1971).

The one best response type

Such a test item consists of a stem (a statement or question) followed by a number of possible answers (usually five). The candidate is asked to mark the one answer that he feels to be best or most appropriate.
An example of such a test item is as follows.

**Example 1**

In diabetic ketoacidotic coma:
A. Body sodium is proportionately more deficient than body water
B. The blood glucose concentration is over 800 mg/100 ml in more than 90 per cent of cases
C. The hyperventilation is stimulated by excessive PCO₂ in the arterial blood
D. Abnormalities found in the examination of the central nervous system characteristically suggest the presence of an upper motor neurone lesion
E. The total intracellular potassium content is abnormally low

Indicate the one best answer with a tick A B C D E

It will be noted that random marking of one of the answers will give a one in five chance of being correct and that if a candidate quite definitely knows the correct answer he does not need to know that the others (the distractors) are wrong. In practice nearly all items of this format have one correct answer, the others being unequivocally wrong, and under such circumstances the knowledge can be tested more usefully by using the multiple true-false format (see below).

The one best response type can, however, be used to test judgment: where all the given answers are not only plausible but *possibly* correct and where one is by consensus the most likely.

**Example 2**

A 45-year-old man known to have had diabetes mellitus for 10 years and to require insulin for its control is brought semiconscious into a general practitioner’s surgery which is 2 miles from the nearest district general hospital. His urine is found to contain more than 2 per cent glucose and ketone bodies are present. The skin is dry, respiration is of the air-hunger type, and the reflexes are normal. The general practitioner should take the following action:
A. Give him 50 ml 10 per cent glucose intravenously and send him to hospital
B. Give him 25 U soluble insulin subcutaneously
C. Wash his stomach out with 5 per cent sodium bicarbonate solution, put 100 ml 10 per cent glucose solution into the stomach, and then send him to hospital
D. Send him direct to hospital in a taxi with a note of the findings but without any preliminary treatment
E. Send him to hospital by ambulance, lying prone with his head to one side, accompanied by a note of the findings but with no preliminary treatment

Indicate the one best answer with a tick A B C D E

The matching type

Two slightly different item forms are used for matching-type questions. In the first a list is given of about five conditions or situations that are in some way related to each other. This is followed by a number of statements each of which is related to a different one of the conditions originally listed. The candidate is required to indicate which of the conditions gives the most appropriate to the statement.

**Example 3**

A. Diabetic ketoacidosis
B. Chronic renal failure
C. Hypoglycaemic coma
D. Subdural haematoma
E. Barbiturate poisoning

Item 1. Skin cold and moist, bilateral extensor responses
Item 2. “Air-hunger” with a serum calcium of 7 mg/100 ml
Item 3. Several days of fluctuating level of consciousness, trace of urinary glucose
Item 4. Serum bicarbonate 30 Eq/litre, PCO₂ arterial blood 60 mmHg

In the other type of matching test there are given, say, two conditions, labelled A and B respectively, and then “Both” (labelled C) and “Neither” (labelled D). These are followed by statements that are required to be labelled A, B, C, or D as most appropriate.

**Example 4**

A. Diabetic ketoacidosis
B. Hypoglycaemic coma
C. Both
D. Neither
Tick the one best answer for each item:

Item 1. Glosousia may be found
Item 2. Neck stiffness is a characteristic finding
Item 3. Bilateral extensor responses are frequently found

The requirement to match obviously makes the candidate sort between a number of possibilities, but the same knowledge can also be tested quite readily with the multiple true-false format.

The multiple true-false format or the multiple true-false-don’t know format

This type of test item consists of a statement or question followed by a number of answers, any of which may be correct, that is none, some, or all. An example is as follows.

**Example 5**

In diabetic ketoacidotic coma:
A. Body sodium is proportionately more deficient than body water
B. The blood glucose concentration is over 800 mg/100 ml in more than 90 per cent of cases
C. The hyperventilation is stimulated by excessive PCO₂ in the arterial blood
D. Abnormalities found in the examination of the central nervous system characteristically suggest the presence of an upper motor neurone lesion
E. The total intracellular potassium content is abnormally low

The answer sheet may be in the form:

Remembering that none, some, or all the answers may be correct, indicate those that you think are correct A B C D E

or

Remembering that none, some, or all the answers may be correct, indicate those that you think are correct by a Y, and those that are incorrect by an N. If you do not know, do not put Y or N.

A B C D E

The National Boards use this type, but do not arrange for the candidate to indicate that he does not know. They tend to formulate it differently, by asking the candidate to mark A if answers 1, 2, and 3 only are correct, B if answers 1 and 3 only are correct, or
C if answers 2 and 4 only are correct, D if answer 4 only is correct, or E if all answers are correct. There are limitations to this format, for there are in fact a total of 32 possible combinations of correct answers, not just five.

The multiple true-false item of the first type is the only one at present used in the part I MRCP UK examination. The answer sheet is arranged so that it can be read by machine, but the candidate has to indicate whether he thinks each answer is right or wrong or that he does not know. Most attributes tested by the other item formats can also be tested in the true-false form and the statistical analysis of the items and of each possible answer can be quite refined. It would seem on general grounds that it is much better for a doctor to be allowed to recognize what he does not know some item of knowledge rather than to be forced to guess the answer. It has also been found that the introduction of this possibility has resulted in test items becoming more discriminating between "good" and "bad" candidates than they were previously (Fleming et al, 1974). Furthermore, additional information can be garnered by the examiners about the knowledge possessed by the candidates when the latter are not forced to guess_ (although obviously an informed guess is likely in the long run to be better than nothing).

Although comments have been made about the various formats that are used in medical objective testing, it cannot be too strongly emphasized that neither the format used nor the precise method of statistical analysis of the answers is the most vital aspect of an objective-type examination. The production of suitable examination material is not only the most difficult part of the exercise, but it is by far the most important. Whatever format of question is used the results are consistent and "reproducible"; differences in the technique of calculating the marks can make some change in the ranking order of the candidates which, although not great, may result in a few "borderline" candidates lying on a different side of an arbitrary pass mark. However, in an overall examination in which the multiple choice paper is only one of the methods of testing, this should not make much difference. Similarly, if the multiple choice test is used as a "screen" this variation should make little difference providing the pass mark is low enough to ensure that those who only just pass are highly unlikely to pass the rest of the examination.

The formulation of test material

Setting really good questions, that is questions that are unambiguous and appropriate in content and difficulty, is not an easy matter, yet it is crucial to the whole quality of the examination. It has been found that a rather elaborate mechanism of "processing" is necessary to ensure that the questions as finally set are satisfactory. The precise details of such a procedure differ from one examination to another, but in general the steps are as follows. Questions are submitted by each of the examiners, the total number being in excess of the number required for any one examination. Alternatively, questions may be submitted by examiners as an ongoing exercise and not just for one particular examination. These are then scrutinized by someone skilled in multiple choice testing methods to ensure that they are technically satisfactory, for example in the multiple true-false format the stem or opening statement must contain no indication of the number of correct answers, and all the answers must be mutually independent. Thereafter the questions may be submitted to a panel of experts in the subject matter of the questions. They will confirm the correctness of the answers as submitted and may make minor modifications to the wording. Sometimes these experts may disagree about an answer in which case it would have to be altered or discarded altogether. They might also object that the subject matter is irrelevant and if the objection was sustained the question would be discarded.

After at least this amount of scrutiny the question will be appropriately classified and entered into a question store. Such a store may in the end contain thousands of questions, many of which will have been used previously (sometimes more than once) and analysed by a computer. This might be programmed so that the analysis is performed on the basis of the overall results of each examination in which the questions were used.

The total size of the bank must partly depend upon the rate of use of multiple choice questions. If the bank is too large in relation to the rate of use, the questions may get out of date and have to be discarded before they are called upon. On the other hand, a small bank raises problems of security or of the possibility that candidates re-sitting the examination may remember a question occurring in an earlier examination (this has been tested and such recognition of previously encountered questions is in fact surprisingly uncommon). The advantages lie strongly with a large question store which is frequently used and which is maintained in an up-to-date state by a large number of experts.

The use of multiple choice examination papers

Examinations composed of multiple choice questions are an excellent means of ranking the candidates with a high degree of consistency. However, 'it is clearly impossible to prescribe a pass mark by statute in the way this is done for orthodox examinations, where the mark awarded by the examiner is subjectively related by the examiner to what he knows to be the pass mark. Thus, when multiple choice papers are first introduced, the marks need to be compared in some way with "orthodox" results so that a pass mark somewhat comparable in standard may be determined. Once this has been done a similar standard may be transferred from one examination to another by comparing the standards attained in questions common to both. When on each occasion, the number
of students taking an examination is large and drawn from a similar population the standard achieved on a given question remains remarkably constant (Owen et al., 1967).

There are least three uses for a multiple choice type paper that requires a pass mark to be established. First, if it is used as a screen as in the part I MRCP UK (Owen et al., 1967; Fleming et al., 1974), those with results above the pass mark are allowed to proceed to the next stage of the examination and those below are excluded. Under these circumstances, fairness dictates that the pass mark should be at such a level that those who fail would be very unlikely to pass the other parts of the examination if they had been allowed to take them.

Secondly, the multiple choice paper may form only one part of an examination in which the marks obtained in each of the parts are aggregated, the resulting total determining the position of the candidate and whether he has passed or failed. In this case it also becomes necessary to ensure that the overall spread of marks between candidates is approximately the same for each part of the examination.

In most medical examinations where examiners award a mark (such as to essay questions, to vivas, or to clinical examinations) a rather close range is traditionally adopted. It is clear that even when the marks of the multiple choice paper are adjusted so that the pass mark is the same as that conventionally adopted for the other parts of the examination, if (as is invariably the case) the spread of marks is much wider in the multiple choice paper, this part of the examination will have an effect on the overall results much greater than any weighting factor that might have been accorded to it. To prevent this effect steps should be taken to ensure a similar spread of marks for each part of the examination; there are a number of ways in which this can be done.

The third situation (similar to the first in many ways) that requires a pass mark is where the various parts of the examination are used separately to determine a pass or fail. The candidate thus has to pass in all the separate parts of the examination - or in, say, three out of a total of four parts, high marks in one part not being allowed to compensate for low marks in another. This technique is traditionally partially used in most professional medical examinations in that a pass in the clinical part is essential to pass the examination as a whole.

The multiple choice technique can be used effectively for feedback to the individual or for feedback to those in charge of teaching or education as it can reveal lacunae in knowledge quite efficiently. The technique can also be used in comparing different groups of students, either in experiments testing different teaching methods or between different schools (Hubbard and Cowles, 1954; Ricketts et al., 1974). For uses such as these, it is obvious that no pass mark is required.

Limitations of multiple choice question papers

Quite clearly, the efficient practice of medicine or one of its specialties requires many attributes and it must be asked how appropriate a multiple choice examination paper is in these circumstances. It is intellectual characteristics that are being tested and although in principle a good deal more than the first of the taxonomic levels described by Bloom (1956) can be appraised, it becomes difficult in practice to be sure that some multiple choice examinations are testing much more than factual recall. This is partly because it is much more difficult to set the more elaborate type of question required to test higher intellectual levels, and partly because the mental process used by one candidate to determine the answer might be different from that used by another. What might have been worked out by one from an understanding of the principles involved may be a matter of factual recall by another. Nevertheless, considerable experience with the use of multiple choice examinations in medicine suggests that a candidate is unlikely to have a good overall performance unless he performs well in the multiple choice paper, and people who do badly in this type of test are very unlikely to be excellent in other types of test. In essence, a man who does not know many of the facts of medicine is unlikely to be good at a whole range of medical functions for so many of which a reasonable store of factual knowledge is essential.

An objective type multiple choice examination paper is a very good instrument for testing a very limited but vitally important range of those attributes that are necessary to make a satisfactory doctor. It would thus be quite wrong for any qualifying examination to consist solely of multiple choice papers. However, if we accept that our aim in an examination is to test each desirable attribute by the best possible means so that we might form a professional profile of each examinee, the techniques of the objective-type examination are highly appropriate and efficient in their own field. It is regrettable that equally effective methods for assessing other attributes of character have yet to be developed, although considerable efforts are being devoted to this end.

Cox, R. (1967) Universities Quarterly, 21, 292
Hubbard, J. P. (1974) British Journal of Hospital Medicine, 8, 209
Hubbard, J. P. (1974) British Journal of Hospital Medicine, 8, 209