Paper published in Psychology Software News (1996), 7, 4-8.

Guidelines for the use of multiple choice and computer presented tests for university assessment

Helen Snow, Andrew Monk and Peter Thompson,
Department of Psychology,
University of York,
Heslington, York, YO1 5DD.

INTRODUCTION

The following notes outline a number of points which may be useful to someone who is thinking about using a multiple choice test for university assessment. They are based on a mixture of established research (see references at end) and our own experience. They are not intended to be a comprehensive reference.

Contents:

SECTION 1: WRITING MULTIPLE CHOICE QUESTIONS

First some terms need to be defined. A multiple choice question can be described as being made up of three parts: (i) the stem in which the body of the question is presented and any necessary information given; (ii) the correct response, and (iii) the distractors, the incorrect responses. Each part will be considered in turn.

The stem

The simplest type of stem asks only for the recall of information. However it is possible to write multiple choice questions which test the ability of students to understand concepts, to interpret information and to apply principles. This

may require longer and more complicated stems. It is also possible to include diagrams, graphs, tables and photographs for analysis. Ideally the information in the stem should not be presented in the same way as it was introduced at the teaching stage. This requires considerable effort from the question setter.

A long stem may penalise people with reading difficulties or English as a second language but it may be possible to base more than one question on a long stem. Avoid including unnecessary information in the stem; the test is not supposed to be a learning exercise. However there is nothing wrong with putting less able students "off the scent" by introducing information linked with one of the distractors.

The correct response

This need not be an absolute truth. For example, the phrase "which of the following is the most likely?" may be used.

The distractors

The vital element of the distractors is that they must be plausible. The more similar distractors are to the correct response and to each other, the more difficult the item is. One possibility is to create distractors based on common errors made by students.

Usually four or five options are presented per item, depending on the number of suitable alternatives to the correct answer which can be found. The greater the number of options displayed, the lower the probability of a student choosing the correct answer by guessing. (This may not be a problem; see later section on guessing.) Obviously if an question contains six options but only two plausible ones then there are really only four options to guess from.

Unintentional clues to the correct answer

These are errors which are often made by test writers. A test-wise student may pick up on them and obtain correct answers by simple elimination of implausible options, with no need for knowledge of the topic being tested.

Some of the fundamental mistakes are listed below:

- The consistent use of words such as "all", "never" and "always" in distractors and "usually", "sometimes" etc. in the correct answers.
- The tendency to make correct answers longer than distractors. This usually occurs because the writer is determined that the answer will be correct beyond doubt.
- Grammatical inconsistencies between stem and distractors. An example is given below:

  Sample Visual Perception Question
  A snowball looks white no matter how much light it is reflecting. This is because of:
  A - the luminosity of snow
  B - lightness constancy

C - the retina contains rods and cones
D - rods are not colour selective

Neither C nor D are grammatically consistent with the stem. The student is likely to eliminate them as a result.

- The use of the option "all of the above" only when it is the correct response.
- The correct response contains a word repeated from the stem but the distractors don't.
- The position of correct responses follows a pattern (e.g. always Bs and Cs). Correct responses should be positioned randomly.

Content

Clearly the question must be relevant to the subject area being tested. It is tempting to make a test more difficult by basing items on obscure details but this is not really measuring students' understanding of the topic. If possible items should be checked by another individual with knowledge of the subject area; in this way mistakes may be highlighted, poor questions modified, and irrelevant questions eliminated altogether.

Order of questions

Items may be arranged randomly or organised in some way for test delivery. It is common to arrange the items in order of difficulty, and / or in groups according to the subject matter.

Number of questions asked and time given to answer

It is difficult to advise on how many questions to present in a single test , and how much time to allow for test completion. Some questions will require one or more minutes to answer; simple recall type items may not require so long. It is advisable to allow enough time for all students to be able to complete the test. There are two reasons for this:

1) slow readers are not discriminated against

2) students are less likely to panic and guess a lot.

As a rough guide, the Visual Perception test used at York allows the students one hour to answer 100 questions that have stems of one or two sentences in length. The Statistics test allows students one hour to answer 14 items each with stems of five or six sentences in length. Each item involves answering a number of embedded questions (3 to 8 questions in each).

Guessing

Guessing is not necessarily a problem. A test is designed to discriminate between candidates, not to provide some absolute score (see Section 2). Scores usually end up being scaled to make the distribution equivalent to other forms of assessment. Scaling can cope with the overall inflation of scores due

to correct guesses. Thus guessing is only a problem in so far as it inflates the measurement error and so reduces reliability.

Scoring

There are a number of possible options open to the test writer. Where there are five or more alternatives to choose from per question the error variance introduced by guessing should not be a problem. If it is only possible to construct one or two plausible distractors one may need to discourage guessing by penalising wrong answers. For example, candidates can be instructed that one point is added to their score for a correct answer and two marks subtracted for a wrong answer. If they are unsure they should then answer "don't know" or skip the question.

This scheme can theoretically produce negative scores. This is not a problem as the scores will almost inevitably have to be transformed with a look up table anyway (see Section 2). It is a mistake to use penalties of this kind as a way of adjusting the distribution of scores to make it comparable with other assessments (again see Section 2).

Obviously whatever scoring method is chosen, it is important to make quite clear at the beginning of the test exactly how items are being marked.

Post-test reviewing of items

The test may be refined after test delivery; an item analysis reveals for each question: its difficulty; its discriminating power, and which distractors were most and least favoured. In this way it is possible to weed out questions which do not discriminate between high and low performers. Distractors which are seldom chosen by students may either be removed or replaced by more plausible options.

Conclusion

Writing a good multiple choice question (i.e. one where the problem is clearly stated, a reasonable amount of thought / calculation is required, and the correct response is not obvious) is deceptively difficult. However the time taken to write the test may be weighed against the time saved when it comes to marking.

SECTION 2. ASSIGNING MARKS TO THE SCORES FROM MULTIPLE CHOICE TESTS

What assessment is supposed to achieve

When we mark an individual piece of assessment from an individual student we are applying *rating* scale. We rate it as a "good 2.1" , a "border line 2.2" etc. To make it possible to summarise the large number of ratings of this kind a

student will accumulate over the years we express that rating as a number. This is sometimes called a "percentage". This is extremely confusing. It is important to be clear that there is no sense in which a student getting a mark of 50% has done half of anything or that a student getting 75% has done half as much again. They are just ratings expressed using a particular numerical scale.

There are two procedural questions to be addressed: (i) how to assign marks to individual pieces of work and (ii) how to summarise those marks and set criteria for assigning final degrees. This paper is only concerned with the former question where the issues are the standard psychometric criteria of: reliability (how much error variance or noise is there in the marks); validity (does the test assess what we want to know about the student) and standardisation (are marks comparable across courses, across years and across institutions).

Pros and cons

One can distinguish two ways of obtaining marks we will call them subjective and objective assessment. What defines the latter category is that there is a mechanical marking scheme that awards points for responses that are either there or not there. An archetypal objective assessment is a multiple choice questionnaire. Another example would be statistics exercises where students have to do calculations. An archetypal subjective assessment is a 4-page essay. To some the terms subjective and objective may seem loaded but as psychologists we know that all behavioural data is suspect and that when treated appropriately subjective ratings are as good as anything else. They will be compared in terms of the issues of validity, reliability and standardisation.

Validity

The advantage of subjective assessment, in the context of degree courses, is that the questions can be very under specified. The words on the paper give very little away about what is required. Definitions of what constitutes a first class answer, a 2.1 answer etc. may be provided for candidates in advance of the exam. However, these will not make explicit precisely what the examiner is looking for in an answer to a particular question. To a large extent what the examiners are marking is the ability of the student to guess what they want and give evidence of the values we all feel are fundamental to university education: clarity of expression, critical thinking and so on. Objective assessment on the other hand has to make explicit what responses are permissible and this makes it impossible to assess these abstract qualities.

This is a different point from the question of whether an assessment "stretches" candidates. Indeed, in our opinion, it is easier to devise objectively assessed questions that test problem solving skills and the ability of a student to apply their knowledge in novel ways. There is always a danger that essay questions will simply encourage the regurgitation of facts and opinions.

Reliability

The disadvantage of subjective assessment is that because the questions are under specified so are the criteria for assigning marks and this can make them inherently unreliable. As psychologists we know that markers will be affected by contrast effects (the quality of the scripts read immediately before this one), fatigue, quality of hand writing and so on. Double marking improves reliability considerably by acting as a filter for aberrations of these kinds. For practised markers, examining boards feel that reliability is acceptable, especially when they average over a large number of pieces of assessment.

Objective assessments should be very reliable certainly, given the same answer, one should always obtain the same score. There is the possibility of errors in the marking but the loss in reliability they induce will be very small compared with the potential differences of opinion in a subjective assessment. A broader view of reliability includes guessing as a potential source of error variance but again, if sensible precautions are taken (see above) this should be small.

Standardisation

It seems that most academics are happy with the comparability of marks achieved by subjective assessment. Objectively marked assessments require scaling (see below) and this makes comparisons across courses very difficult. Both types of mark are of course amenable to mechanical standardisation procedures.

Objectively marked assessments are generally more practical to reuse from year to year and so, potentially, provide more comparable marks in that sense.

Conclusions

Objective and subjective assessment both have their pros and cons. In terms of validity, they tend to get at different things, but these are all things that we want to assess. Objective assessment wins on reliability but looses on standardisation. A reasonable conclusion would be that an undergraduate degree should contain both.

Turning test scores into marks

The main problem faced by someone using a multiple choice questionnaire (once they have made up the questions!) is how to convert the test scores into marks i.e., points on a standard rating scale where 50 is a borderline 2.2 and so on. Probably the best way to do this is with a look up table. An example is given below.

Lowest test score to get	a mark of	
0	0	
25	10	Fail
40	25	
55	35	Pass
60	42	
63	45	Third
66	48	
70	52	
74	55	2.2
78	58	
81	62	
83	65	2.1
85	68	
87	75	
89	80	First
91	85	

The table was used to give marks to students for their course work in Data Analysis here. It works as follows. Any student with a score of less than 25 gets 0, a student with a score >= 25 and <40 gets 10, and so on up to candidates who get 91 or more who all get 85. This look up table can be used "by hand" or, when using a spreadsheet, with a "lookup" function (note that Excel requires that values are in ascending order, as here). The mapping provided by this table (while monotonic!) is very flexible. There are 10 score points in the third category, 11 in the 2.2 category, 6 in the 2.1 category and 4 in the firsts.

A look up table of this kind is constructed by deciding what test score corresponds to each class boundary and then assigning the 16 rating scale points accordingly. This may seem arbitrary but it is essentially what one does every time one marks an essay. The difference here is that it was done once only and then applied mechanically to all the objective test scores.

The alternative is to devise some arithmetic transformation. For example, one might subtract 15 from the percent correct to make the maximum score 85. Generally, however, simple measures like this give very peculiar distributions. Even quite complex non-linear arithmetical transformations are not guaranteed to deliver a sensible distribution of marks. Given what we are doing is assigning a rating based on our knowledge of the nature of test and the score of the student, which is more arbitrary: an arithmetic transformation that happens to be convenient (if you can find one), or, a considered judgement of what each range of test scores should mean?

It can be argued that one is "throwing away precision" by using a discrete scale (there can be no student with mark of 60 for example). However, it can equally well be argued that such precision is illusory. It is certainly unnecessary when several marks are summarised to award a degree class. Most importantly, these marks are directly comparable with marks obtained in subjective assessments.

SECTION 3: COMPUTERISED TESTS

It is possible to present a multiple choice test on a computer as opposed to the usual pen-and-paper format. A number of software packages are available for the creation and delivery of computerised tests. This section is based on our experience of Question Mark, a package that allows eight different types of objective-type question to be presented (among them multiple choice, multiple response, matching / ranking and fill in the blanks). It also marks the students responses as they take the test and presents summaries in a number of electronic formats (e.g. Excel and SPSS) at the end of the test.

Another possibility is computer marking tests presented on paper. There are two technologies available to do this. The older one is optical mark recognition (OMR). The alternative is optical scanning (OS). In OMR what is recognised is a simple mark (e.g. tick) at a particular position on a form. This has the disadvantage of requiring accurate printing and limits the form of response. OS recognises the whole form, printing tolerances are less strict and numerical or written responses can be recognised. The price paid for this flexibility is that OS is slower but this should not be a problem with the numbers involved for university assessment.

The advantages and disadvantages of computer presentation and computer marking are discussed below.

Computerised tests: advantages:

Versatility of presentation

Computer presentation is a more versatile method of test presentation than printed versions; one can display as many pictures, diagrams and colour photographs as one likes without the extra printing costs. One can even display video and audio clips.

Speed and efficiency of marking

The main benefit of the computerised test is that it is automatically marked, and the data may then be transferred onto a spreadsheet with no manual data input required.

Computerised tests: disadvantages:

Resources

There are likely to be more students than there are computers available for computer presented tests, furthermore they will not be set out in the room as one would want in an examination.

These practical problems can be surmounted. A computer presented examination for the Visual Perception module was recently sat in this department by 102 candidates. Because of the limited number of computers available it was run three times. To ensure that the three different groups of students did not communicate with each other the second and third groups had to wait in a separate room while the first group was finishing the test and leaving the building. The second group then sat the test while the third group continued to wait. Finally, the third group sat the test, after only just over an hour's wait. Most students felt that this small inconvenience was more than ably compensated for by the release of the marks less than 24 hours later.

Security

To ensure against copying from another candidate the size of the text on a computer terminal must be made as small as is reasonable. One benefit of using Question Mark is that it will present the responses in a different random order for each candidate. This makes copying the answers of the person in front much more difficult. As in any other examination invigilators provide the main precaution against cheating.

Computerised assessment involves the storing of both test questions and students' marks on computer, though now that most people use word processors and spreadsheets this will generally apply to all forms of examination. The issues this raises are:

i) the provision of back ups to prevent accidental erasure or hardware failure, and

ii) measures to prevent malicious damage and cheating.

Question Mark safeguards against hackers by storing both question and answer files on disk in encrypted form so that the contents of the files cannot be examined without using Question Mark software. As a further precaution, passwords may be assigned to files.

Finally, although Question Mark records each student's answer to every question, this record does not have the face validity of a paper answer with the student's handwriting on it. This makes it difficult to prove beyond doubt that a person gave a particular answer to a question and that the computer did not make an error recording the responses.

Conclusion

Computerised tests are a very efficient way of assessing students; not a single sheet of paper is used and test marks are available almost as soon as the exam is completed. Clearly the savings in time and effort when it comes to marking are tremendous. However, there are practical problems to surmount and substantial planning is required from the outset.

BIBLIOGRAPY

Brown, G. & Pendlebury, M. (1992). Assessing active learning. CVCP / USDU, Sheffield.

Bull, J. (1993). Using technology to assess student learning. TLTP Project ALTER. 1993 ISBN 1-85889-091-8.

Heywood, J. (1977). Assessment in higher education. Wiley, London.

Mathews, J. (1981). The use of objective tests. Revised Ed. Lancaster: University of Lancaster Press.

Thorndike, R.L. & Hagen, E.P. (1969). Measurement and evaluation in psychology and education. 3rd Ed. Wiley, New York.

ACKNOWLEDGEMENTS

Thanks are due to: Rob Stone for his various contributions; John Nash for answering our questions about the administration of examinations, and to the University of York Teaching Initiative for funding Helen Snow's work on this project.