Llywodraeth Cynulliad Cymru
Welsh Assembly Government

www.cymru.gov.uk

Multiple-choice testing as an assessment tool

Guidance

Date of issue: September 2010

Multiple-choice testing as an assessment tool

Audience

Awarding bodies involved in the design, development and assessment of multiple-choice testing.

Overview

This document provides guidance to awarding bodies on the use of multiple-choice questions as an assessment tool in order to improve reliability and validity of this process.

Action required

None – for information.

Further information

Any questions should be addressed to:
Pamela Gay
Qualifications and Learning Division
Department for Children, Education, Lifelong Learning and Skills
Welsh Assembly Government
Tŷ'r Afon
Bedwas Road
Bedwas
Caerphilly
CF83 8WT
Tel: 01443 663622
e-mail: Info.quals@wales.gsi.gov.uk

Additional copies

This document can be accessed from the Welsh Assembly Government website at www.wales.gov.uk/publications

Related documents

None.

ISBN 978 0 7504 5678 4

F2901011

Contents

GUIDANCE ON MULTIPLE-CHOICE TESTING

Multiple-choice tests are used for a very wide variety of qualifications. As with any method of testing or assessment, the onus must be on the body responsible for awarding the qualification to ensure that the method of testing/assessment employed, is appropriate and this must include taking account of the framework within which the qualification sits.

1. What are multiple-choice tests?

Multiple-choice tests are tests consisting of questions with a series of alternative answers from which learners select the one which they believe to be correct.

Probably the most common form (although there are other types) of multiple-choice question, is the question which has four alternative answers, only one of which is correct.

2. Basic terminology used

The question is called '*the stem'.*

The incorrect options provided are called '*the distracters'.*

The correct option provided is called '*the key'.*

So for example:

stem	*In which European City will the 2012 Olympic Games be held?*
distracter	a. Barcelona.
key	b. London.
distracter	c. Paris.
distracter	d. Coventry.

The question and the alternative answers are known as *'the item'.*

As with any other assessment tool or test the multiple-choice test must be *reliable - it* must consistently measure the same thing. It is important not to fall into the trap of thinking (because of the use of the word reliable in common speech) that if a multiple-choice test or indeed any other type of test is reliable it is good. It might just consistently be measuring the wrong thing!

Also as with any other test a multiple-choice test must be *valid* – it must test what we require and expect it to test.

Item banks are literally a store of items which have been used and are considered to be valid and reliable, they have in effect been 'tested' or piloted. They are therefore banked to be used in the construction of future tests. Not surprisingly they do need to be regularly reviewed and updated as necessary. (See section 10 on Testing the test).

3. Purpose

Multiple-choice tests like other types of assessment can have several purposes. Within a credit framework, it could be used as a formative assessment tool, allowing an opportunity to provide feedback to a learner, but it can also be used within the summative assessment process.

4. What needs to be considered before deciding to use multiple-choice tests?

As with most things multiple-choice testing is unlikely to be the panacea for all issues in regard to assessment, as would be expected it needs to be considered carefully. Section 6 on Advantages and disadvantages may help to inform and listed below are some other points worth consideration (**N.B.** this is not an exhaustive list):

- The development of multiple-choice tests, usually involves quite a large 'up-front' investment. This in many cases is recouped later, but that will not always be the case.
- Writing good multiple-choice questions is not as easy as it appears. It requires expertise in writing multiple-choice questions and usually subject expertise How will security of the test be dealt with?
- How will topicality of the questions be ensured – this may relate to something as simple as a date has passed for example: Where will the 2012 Olympics be held? or it may relate to changes of legislation.
- How will you ensure there are sufficient numbers of questions in the question bank to ensure that anyone having to re-sit does not get exactly the same questions again, that the test detail is not passed on from one group of learners to another etc?
- How will you ensure the questions are valid and reliable?
- How will you set the pass mark?
- Will you need to establish a policy in regard to revision and maintenance of questions?
- Will you need to establish a policy in regard to evaluation of the questions and/or the full tests?

5. Multiple-choice test formats

5.1 Multiple-choice test question

This is the simplest format and probably the most common.

There is a question *(stem),* four alternative answers including one correct answer (*key)* and three incorrect answers (distracters).

For example:

Which famous impressionist artist painted the Water Lilies?

a. Monet.
b. Gauguin.
c. Manet.
d. Renoir.

(Correct answer: a)

5.2 Extended matching question

This format which is used by the UMAP tests (see section below), gives a list of possible answers, and then a number of different situations are described, for which one of the answers on the initial list is the correct one.

An example of this would be that a list of ten pharmacological preparations is given and descriptions of symptoms exhibited by five patients. The learner then has to select the correct preparation for each patient. Each preparation can be selected more than once or not at all.

'If there are few plausible distracters and a lot of homogenous items then matching questions are better than simple multiple-choice items[1].'

5.3 Data sufficiency

This tests the quantitative reasoning ability using an unusual set of directions. It is a unique type of maths question created especially for the Graduate Management Admission Test (GMAT) used widely in the USA.

The examinee is given a question with two associated statements that provide information that might be useful in answering the question. The examinee must then determine whether either statement alone is sufficient to answer the question; whether both are needed to answer the question or whether there is not enough information given to answer the question.

5.4 [2]Examples of multiple-choice Items:

5.4.1 Single correct answer –

In items of the single-correct-answer variety, all but one of the alternatives is incorrect.

[1] Measurement and Assessment in Teaching by MD Miller, RL Linn and NE Gronlund 10th Edition 2008 Prentice Hall.

[2] How to prepare Better Multiple-choice Test Items: Guidelines for University Faculty by Steven J Burton, Richard R Sudweeks, Paul F. Merrill and Bud Wood 1991.

The learner is required to identify the correct answer.

Example:

Which of the following is a European city?

a. Melbourne.

b. Bangkok.

c. Rome.

d. Watford.

(Correct answer: c)

5.4.2 Best answer –

In items of the best-answer variety, the alternatives differ in their degree of ‘correctness’. Some may be completely incorrect and some correct, but one is clearly more correct than the others. This best alternative serves as the answer, while the others are the distracters.

The learner is required to identify the best answer.

Example:

Monopolies can cause problems in a market system because they:

a. Create external costs and imperfect information.

b. Lead to higher prices and under production.

c. Make such large profits.

d. Manufacture products of poor quality.

(Correct answer: b)

5.4.3 Multiple response –

In items of multiple response variety, two or more of the alternatives are correct answers whilst the others are distracters.

The learner is required to identify each correct answer.

Example:

Which of the following is a characteristic of a virus?

a. It can cause disease.
b. It can reproduce by itself.
c. It is composed of large living cells.
d. It lives in plant and animal cells.

(correct answers: a & d)

5.5 [3]Question formats

This information is taken from the Praxis exams, which are educational tests that American states use as part of their teacher certification process.

There are three types of questions used in the Praxis exam. Each type of question represents an increasing level of complexity. The three types of question appear randomly on the test.

5.5.1 Type A: Basic knowledge

Type A assesses basic knowledge. These questions are factual and simply require the knowledge of a piece of information.

Example:

In which year did the Titanic sink?

a. 1915
b. 1944
c. 1912
d. 1899

(correct answer: c)

5.5.2 Type B: Application

Type B questions are designed to test basic knowledge and to use it in context. These questions require application of information in a specific context. Such questions do not require the learner to address the full complexity of real life situations, but they demand more than simple memorisation of facts.

[3] Understanding Multiple-choice PE praxis.com.

Example:

James, at the age of two months, is very active and wriggles frequently. The findings of a study on the origins of temperamental or constitutional personality differences would predict that:

a. James will be very quiet and docile by age five.

b. James will succeed in school.

c. James will very likely be active and unable to sit still for long as a small child.

d. James will be neurotic.

(correct answer: c)

Example:

Which of the following concepts would explain why it is easier to maintain balance during a headstand than during a handstand?

a. The height of the centre of gravity is lower in the headstand than in the handstand.

b. The line of gravity is over the base of support in the headstand, and outside the base of support in the handstand.

c. The length of the moment arm is longer in the headstand than in the handstand.

d. The magnitude and direction of force are greater in the headstand than in the handstand.

e. The frictional forces are greater in the headstand than in the handstand.

(correct answer: a)

5.5.3 Type C: Analysis, synthesis and evaluation

Type C questions require that learners analyse, synthesise, evaluate and make a decision. This type of question is based upon a hypothetical situation and asks the learner to use their knowledge in order to make judgments. These questions often involve a scenario and require integration of knowledge and decision making. Sometimes the learner must decide the most appropriate steps to take, given a hypothetical case or situation.

Example:

A student suffers an injured ankle while running in a rounders match. The teacher questions the student about how the injury occurred and about the area affected. The teacher examines the indicated area. The symptoms are typical of a sprained ankle, although the injury may in fact be more severe. Which of the following steps should be included in the first aid administered to the student?

i. Elevate the injured leg.
ii. Apply ice to the injured area.
iii. Apply direct pressure to the site of the injury.

a. i only.
b. ii only.
c. i & ii only.
d. i & iii only.

(correct answer: c)

6. Advantages and disadvantages of multiple-choice testing

It can be argued that every form of assessment has advantages and disadvantages, and multiple-choice testing is no exception. Listed below are a number of advantages and disadvantages, it is not given as an exhaustive list and it is recognised that there are many factors which may influence the opinion regarding advantage or disadvantage.

6.1 Advantages

- They are often quicker to administer than other tests particular ones which require written responses.
- They lend themselves to on-line assessment.
- Many questions can be asked in a relatively short period of time because learners are not required to write large amounts of material.
- Because many questions can be asked, this lends itself to sample across a whole unit/qualification. This forces learners to study the whole rather than speculate on what will be asked in a particular paper.
- Marking is fast and consistent and is considered to be thoroughly objective.
- Marking can be ‘computerised’ which can also then provide statistical information that can be used in the test evaluation process.

6.2 Disadvantages

- Public perception can be that this is only suitable for lower level qualifications and that it is possible to guess the correct answers.
- It cannot assess the direct application of practical skills.
- Writing good quality multiple-choice questions is demanding and time consuming.
- It cannot assess learners abilities to work at problems or projects over a protracted period of time.
- When writing multiple-choice questions, there is absolutely no room for error. If for example the correct answer is misspelt, then arguably the learner cannot give an answer.

7. Security of multiple-choice tests

There is a security aspect to all tests and exams and this includes and is no different when considering multiple-choice tests.

It will be advisable to have clear policies and procedures in place and there is no reason why they need to differ in relation to the style of exam i.e. essay or multiple-choice. It may be appropriate to clearly state that the policies and procedures do relate to **all** types of exams/tests offered.

Where there are likely to be differences in requirement, is when an exam/test is computer based and/or on-line. It is highly likely that there will be a need to establish different policies and procedures for computer based/on-line and paper based systems.

8. Developing multiple-choice questions

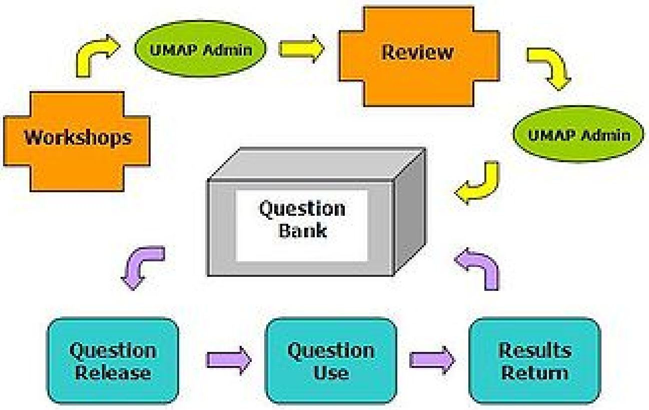
It is true to say that developing good quality multiple-choice questions is not an easy or a quick task. It does require expertise in writing multiple-choice questions but it also requires expertise in regard to the subject matter.

The Universities Medical Assessment Partnership (UMAP) provides an interesting case study.

8.1 Case study on developing multiple-choice tests – UMAP

The UMAP project was originally funded through the HEFCE Fund for the Development of Teaching and Learning, Phase 4. Funding was awarded for a three year set up term which ran from January 2003 to December 2005 http://en.wikipedia.org/wiki/Universities_medical_assessment_partnership - cite_note-4#cite_note-4. At the end of the funded period the original five partners agreed that UMAP would move to become self-funded. At this time the consortium invited other UK medical schools to join.

The UMAP process



UMAP is a collaborative project where each member school contributes equal effort and resource to the development of a written assessment item bank dedicated to multiple-choice questions and extending matching questions items. Each partner hosts question writing training workshops, facilitated by the central project team. Invited item authors are members of either NHS or university staff. Items are brought forward into the item bank in draft form before being allocated for editing and approval by one of UMAP's many question quality assurance teams. Items are fed back into the bank with those being approved where they are ready for use in any notified high stakes examination taking place at a partner medical school. Partner schools select a range of items from the UMAP bank in advance of an examination and scrutinise these in order to achieve the correct mix of items for the intended examination. Anonymous examination results are later fed back to UMAP so that item analysis can be undertaken, removing any items from further use if necessary. Feedback is then returned to the original authors including the proportion of correct student answers per each item usage instance.

The test is a test of what is regarded as “core knowledge” for doctors. The success of UMAP is felt to lie in the way in which the tests are designed, as in the diagram above, in which many experts are involved; the test item development and item review processes are separated. The UMAP tests are not the only form of assessment used. In addition doctors are assessed through Objective Structured Clinical Examinations and mini-clinical examinations in which their performance in dealing with actual patients or simulations is observed and assessed. They are also assessed through portfolios. The UMAP bank of items now holds about 5000 items of which 125 will be used in any one test.

9. Blooms taxonomy

Bloom's taxonomy classifies the types of knowledge that might be required by test papers into six levels, each of which is more demanding than the last.

It can be useful to classify items according to the taxonomy and to ask item writers to try to write items that work down the taxonomy as a way of increasing the intellectual demand of the items. Although this can be a useful exercise, it is not the whole answer as it does tend to prompt non-productive arguments about the levels that should be assigned to different items, and it also turns the focus from the requirements of the syllabus or unit/qualification content (which should represent what needs to be tested) to the merits of the taxonomy itself.

However, the taxonomy does illustrate how it is possible to consider more than just the learners' memories when setting questions, so it does have value. The main levels of the taxonomy are listed below in increasing order of supposed demand from the top down[4].

Category **(Name of level)**	**Example and key words**
Knowledge: Recall data or information.	**Examples:** Recite a policy. Quote prices from memory to a customer. Knows the safety rules. **Key words:** Defines, describes, identifies, knows, labels, lists, matches, names, outlines, recalls, recognises, reproduces, selects, states.
Comprehension: Understand the meaning, translation, interpolation and interpretation of instructions and problems. State a problem in one's own words.	**Examples:** Rewrites the principles of test writing. Explains in one's own words the steps for performing a complex task. Translates an equation into a computer spreadsheet. **Key words:** Comprehends, converts, defends, distinguishes, estimates, explains, extends, generalises, gives examples, infers, interprets, paraphrases, predicts, rewrites, summarises, translates.
Application: Use a concept in a new situation or unprompted use of an abstraction. Applies what was learned in the 'classroom' into novel situations in the workplace.	**Examples:** Use a manual to calculate employee's holiday allowance. Apply laws of statistics to evaluate the reliability of a written test. **Key words:** Applies, changes, computes, constructs, demonstrates, discovers, manipulates, modifies, operates, predicts, prepares, produces, relates, shows, solves, uses.
Analysis: Separates material or concepts into component parts so that its organisational structure may be understood. Distinguishes between facts and inferences.	**Examples:** Troubleshoot a piece of equipment by using logical deduction. Recognises logical fallacies in reasoning. Gathers information from a department and selects the required tasks for training. **Key words:** Analyses, breaks down, compares, contrasts, diagrams, deconstructs, differentiates, discriminates, distinguishes, identifies, illustrates, infers, outlines, relates, selects, separates.

[4] Workbook on Multiple-choice Testing Norman Gealy 2004.

Category **(Name of level)**	**Example and key words**
Synthesis: Builds a structure or pattern from diverse elements. Put parts together to form a whole, with emphasis on creating a new meaning or structure.	**Examples:** Write a company operations or process manual. Design a machine to perform a specific task. Integrates training from several sources to solve a problem. **Key words:** Categorises, combines, complies, composes, creates, devises, designs, explains, generates, modifies, organises, plans, rearranges, reconstructs, relates, reorganises, revises, rewrites, summarises, writes.
Evaluation: Make judgments about the value of ideas or materials.	**Examples:** Select the most effective solution. Hire the most qualified candidate. Explain and justify a new budget. **Key words:** Appraises compares, concludes, contrasts, criticises, defends, describes, discriminates, evaluates, explains, interprets, justifies, relates, summarises, supports.

For more information on Bloom's Taxonomy go to:
www.sos.net/~**donclark**/hrd/**bloom**.html

10. Testing the test

In order to ensure the test items do what they are required to, it is usually expected that items will be 'tested' or piloted. This can be completed in several ways, for example:

- new or untested items are included in a multiple-choice test simply to test that item, and they will not form part of the learners results on that test; and
- a test is completely compiled of new items and is then 'tested' on a group of learners who have already been assessed in some other way.

Obtaining statistics relating to multiple-choice tests is usually considered to be relatively easy. The parameters defining a test item are discrimination (D) and facility (F).

Discrimination (D) compares the number of correct responses to an item for the upper and lower 27% of the learners being tested (based on test score). If for any one item the number of correct responses to an item from learners in the lower 27% of the whole test is greater than the number from those in the top 27%, then that particular item may not be effectively discriminating between learners.

Facility (F) is the percentage of learners being tested obtaining the correct answer. Generally:

- if F < 30% the question is hard;
- if F is between 30 and 75% the question is satisfactory; and
- if F > 75% the question is easy.

The decision of whether or not an item is acceptable in relation to difficulty invariably must be a judgment made by those setting the test. Inclusion of hard items with high discrimination maybe useful to rank learners.

Items with acceptable discrimination and facility can be stored in the item bank for future use.

11. Setting the pass mark

It is inappropriate for a pass mark of a test to be set arbitrarily, it should be justified with empirical data.

The Angoff Method is a widely used standard setting approach to test development. It is basically a study that test developers use to determine the passing percentage (cutscore) for a test.

The Angoff Method relies on subject matter experts who examine the content of each test question and then predict how many minimally qualified learners would answer the question correctly. The average of the judges' predictions for a test question becomes its predicted difficulty. The sum of the predicted difficulty values for each item averaged across the judges and items on a test is the recommended Angoff cut score. Here is a real world example that illustrates the process:

Let's say a test developer needed to determine the passing grade for a language exam that tested a person's ability to read Arabic. Using the Angoff Method, the developer would employ a number of subject matter experts (in this case, Arabic-language experts) and ensure that they were properly trained on how to use the Angoff Method, as well as informed on the test's purpose.

The Arabic-Angoff Panel would then rate each test item based on whether or not a minimally-qualified learner would answer the item correctly or incorrectly. Once the first round of ratings had been conducted, everyone on the panel would be given access to the ratings of the other sector matter specialists so that they could compare what they determined about a particular item. Then, the sector matter specialists would be asked to rate the items again for a second round. The second round of rating would give the sector matter specialists the opportunity to review their initial rating of an item and decide whether or not they might like to change their decision based on the expert judgments of the other panelists. This second round of ratings would be averaged across the sector matter specialists to determine the final cutscore for the test.

In the end, using the Angoff Method ensures that the passing grade of a test is determined empirically, which is necessary for a test to be legally defensible and meet the Standards for Educational and Psychological Testing. Other standard-setting methods that testing professionals sometimes employ include the Ebel, Nedelsky, and Bookmark methods[5].

12. Analysis and evaluation

It is accepted and expected that any test or exam will be evaluated, not least to ensure it is both reliable and valid. Most organisations will have systems and procedures in place to for this to be completed on a timely basis and this should include any multiple-choice tests.

The potential advantage with multiple-choice tests in regard to evaluation, is that because they are often ‘marked by computer’, the software package will usually have inbuilt systems which can provide statistical analysis relating to both the test as a whole, and individual items (questions).

Probably the most common statistic provided is that which relates to the number of learners answering an item correctly. This is known as the facility index, and is basically the measure of how difficult learners find an item. It would also be expected for the discrimination index to be calculated. This basically is a measure of how closely scores on the tests as a whole relate to scores on an individual item. The expectation would usually be that, learners who score highly on the test as a whole are more likely to be the ones who answered an individual item correctly. If the discrimination index is positive, then this is the case, if however it is negative the opposite is true. Understandably further investigation would be required in relation to an item giving this result.

Considerable statistics can be gathered in relation to multiple-choice tests and there are many books written on the subject. Other commonly calculated information includes:

- The mean (average) score.
- The standard deviation of the scores.
- The reliability coefficient usually calculated by Kuder Richardson’s formula 20.
- The standard error of measurement.

In completing the evaluation of items/test it also provides the opportunity to check more general aspects such as the ‘topicality’ of items and also the validity of the test. This later point can only really be ensured by having items which cover an appropriate sample of the content of the unit/qualification.

[5] What is the Angoff Method? by Jocelyn Project Director Testing and Training Services ALTA October 2008.

13. Top tips for developing multiple-choice items[6]

i. The stem should be meaningful by itself and should present a definite problem.

ii. The stem should include as much of the item as possible and be free of irrelevant material.

iii. A negatively stated stem should be avoided and only used when significant learning outcomes require it.

iv. All alternatives should be grammatically consistent with each stem.

v. An item should contain only one correct or clearly best answer.

vi. Items used to measure understanding should contain some novelty, but too much novelty should be avoided.

vii. All distracters should be plausible – the purpose of the distracter is to distract the uninformed from the correct answer.

viii. Verbal association between the stem and the correct answer is to be avoided.

ix. The relative lengths of the alternatives should not provide a clue to the answer.

x. The correct answer should appear in each of the alternative positions an approximately equal number of times but in a random order.

xi. Special alternatives like 'all of the above' or 'none of the above' should at most, be used sparingly.

xii. Do not use multiple-choice items when other options are more appropriate for example:

- if there are only two possible responses then True/False is better;
- if there are few plausible distracters and a lot of homogeneous items then matching questions are better;
- for problem solving, short answer questions are better; and
- for complex achievement goals (such as practical activities) performance based assessment tasks are better.

[6] Measurement and Assessment in Teaching by Miller, Linn and Gronlund; 10th Edition 2008, Prentice Hill.