

Formal Testing (double module)

WORKSHOP LEADER'S NOTES

Aims

This double module aims to develop an understanding of testing for the purpose of summative assessment. It is particularly concerned with the development of items which go beyond testing the recall of information. It considers the features of items which test process skills and the understanding of concepts.

Objectives

1. Ability to distinguish between questions which assess ability to recall facts and those which test understanding of ideas through application.
2. Identification of the characteristics of questions which assess process skills and understanding.
3. Knowledge of how to mark open questions.
4. Ability to devise questions to assess specified process skills and understanding.

Note

This module is described as 'double' since it requires at least twice the time of any other module. It could be extended further still if there is the opportunity to extend Task F5 sufficiently to produce items good enough to be tried out with children. Each item has to be thoroughly discussed and revised until it is unambiguous and appropriate to the children. (If it is possible to conduct a school trial, each group of participants might concentrate on refining its two best items for this purpose, giving a test of about 12 items long.)

Task F1

Organisation: Participants in groups of about four.

Materials needed: Task F1, Resource 12.

Time: Allow at least 10 minutes for the task and 20 minutes for plenary feedback.

Instructions: Tell the participants to:

- First answer each of the items.
- Then consider what skill, idea or factual knowledge they were using in answering each one.
- Identify an appropriate age group for each item.

Task F1

Read the three test items in Resource 12 and answer them. Then decide, as a group:

- (i) What skill, idea, or factual knowledge is being assessed by each item?
- (ii) For what age levels are these items suitable? What are the reasons for your suggestions?

Plenary discussion

Ask one group to report on answers to both i) and ii). Then invite comments, not full reports, from other groups.

Points to emphasise

- The difference between recall (item 1) and application (item 2).
- Although item 3 also involves the idea of dissolving, this is used as a context for a question which requires an investigation to be planned. Answering depends on being able to plan a controlled investigation not on the understanding of dissolving.
- Comparing item 1 and items 2 and 3 should lead to a discussion of the pros and cons of multiple-choice formats (see background information).
- Age appropriateness is determined by the presentation and wording of the item as well as the ideas and skills involved.

Task F2

This takes the activity introduced in Task F1 much further. It is an important exercise in analysing test items as a basis for creating items to test process skills and the understanding of ideas.

Organisation: Participants work in pairs for 40 minutes to complete the activity for all the items in Resource 13. Then they move into groups of four to compare responses and identify points of disagreement for a further 15 minutes.

Materials needed: Task F2, Resource 13.

Time: Allow at least 40 minutes for the task in pairs and 15 minutes in groups in preparation for plenary feedback (which may take a further 45 minutes).

Instructions: As for Task F1. This task extends the range of items to be considered and opens up new points for discussion.
Tell the participants to:

- Try to pick out the main skills or ideas that are assessed, not all that are possible. This is particularly relevant in relation to 'observation' and 'communication' which can often be thought to be assessed in all items. These skills should only be listed when they are the main ones.
- Decide whether knowledge or application of the content is being assessed. This is essential since all the items have to be about some content. Suggest that the participants ask themselves: '*Could this be answered without more knowledge of the content than is given?*' If so then it is not testing that knowledge.

Task F2

For each test item in Resource 13, decide what idea(s) and process skills(s) are being assessed and for what age level it is appropriate? Work in pairs and complete a table of the following kind:

Item	Idea(s) assessed	Process skill(s) assessed	Appropriate age
1			
2			
3			
4			
5			
6			
7			
8			

Plenary discussion

Ask the first group to finish. Write their table of answers on an OHP transparency, or a large sheet of paper, or a blackboard, according to what is available. Begin the plenary feedback by asking for additions or changes to these responses. Where something is challenged, ask for the reason and discuss it. Also invite criticisms of the items but at the same time ask for suggestions of how they might be improved.

Points to emphasise

- Some areas of disagreement about what is assessed can probably be sorted out by referring back to the advice given in the instructions, that is to identify the main skill and to consider whether the question requires the child to use conceptual knowledge gained previously in order to answer the question.
- Other disagreements may reveal different understandings of the meaning of certain process skills, since items show the skills in operation. There are genuine differences in views of, for example, the meaning of an 'hypothesis'. It is important to talk them through. Reference to background information for Module D may help.
- What is assessed depends as much on what is given credit in marking as on the question. Thus the marking scheme must reflect the intentions of the item. At the same time the 'wrong' answers have value for the teacher in giving clues to children's understanding and in teacher-marked tests, these should be noted.
- Marking can be seen as a problem where children give open answers. This is a link to Task F3.

Task F3

Organisation: Participants in groups of four to six.

Materials needed: Task F3, Resources 13 and 14.

Time: Allow at least 20 minutes for the two parts of the task and about 15 minutes for plenary feedback.

Instructions: Tell participants to:

- Refer back to item 2 in Resource 13.
- Read the answers which were written by some 10-year-old children.

Task F3

Work as a group. Read the answers (Resource 14) given by some 10-year-old children to item 2 in Resource 13.

(i) Can you mark each of these answers as 'correct' or 'incorrect'?

If not, which do you think deserves most marks? Which deserves least marks? Why?

(ii) Devise a mark scheme for this question.

Plenary discussion

Ask one group to give its answers to sub-task (i). Children's answers are rarely wholly correct or wholly incorrect. Discuss each answer in turn. Provided we can define categories which are sufficiently clear and distinct from each other, graded marks can be given.

Ask groups to give their answers to sub-task (ii) briefly. If no-one suggests a scale, propose and discuss the idea of giving:

2 marks to b);

1 mark to d) and to a);

0 marks to c) (since it does not answer the question).

Points to emphasise

- The mark scheme and the question should be designed together. If a fair scheme cannot be devised, then the question may need to be reworded to indicate more directly what kind of answer is intended. For instance it may well be argued that answer c) *does* answer the question in the eyes of a child. However we can't be certain that the child who gave it had noticed the pattern in the span and number of coins. Thus the link between the mark scheme and question is seen to be crucial and it is important for the two to be considered at the same time.
- It is not necessary to give numerical marks to the different kinds of answer. The answers could be assigned to qualitative categories, given labels A, B, C, etc or whatever. This enables differences between different kinds of answers to be noted without suggesting that some are 'better' than others. It is possible to count the number of answers of the children in each category and to use this information in relation to some overall criteria. When numbers are used and totalled it isn't possible to know whether a score of 10 was made up of three 3s, a 1 and a 0 or five 2s.

Task F4

Before introducing Task F4 it is useful to distribute Resource 15. This gives a brief overview of the way in which items for testing process skills and conceptual understanding are constructed. Allow time for reading and then review the items in Resource 14 to see to what extent they 'fit' the generic descriptions. Making this link will help participants use the descriptions in Resource 15 in this next task.

Organisation: Participants work in pairs.

Materials needed: Task F4, Resource 15.

Time: Allow at least 40 minutes for the task in pairs to produce first drafts of the items.

Instructions: Tell the participants they are to:

- Draft three items, with their mark schemes, according to the specifications.
- Be prepared to defend their approaches and, whilst it is recognised that polished products can't be produced in the time, the items should be ones they consider useful in practice.
- Choose the format (open or multiple-choice) for themselves.
- Write their items on large sheets of paper or on overhead projector transparencies.

Task F4

Devise a test item and mark scheme to assess each of the following specifications:

- (i) An item using the context of the conditions needed by plants for life and growth and testing the process of controlling variables, suitable for 8/9 year-olds.
- (ii) An item applying simple ideas about forces suitable for 7/8 year-olds.
- (iii) An item using the context of shadows and testing the process skill of predicting, suitable for 9/10 year-olds.

Plenary discussion

Discuss each question and mark scheme and criticise constructively with a view to writing better second drafts. This may be done as part of the next task in this module.

Points to emphasise

- Avoid questions which depend both on process skills and the use of prior knowledge. It will not be possible to know if failure is due to lack of the skill or the idea or both. This may not matter if an overall score is computed, but it is more helpful to have sub-scores for skills and understanding of ideas.
- In general a single question is not enough to test a process skill, since its use is influenced by the content (even if knowledge of this content is not required). Ideally several questions, spread across different contexts should be used for each skill.
- Avoid gender bias in selecting the content and context (ie situations commonly associated mainly with the activities of either boys or girls but not both).
- Avoid questions that could be answered by recall without either applying concepts or using process skills.
- It is useful to use the same context or subject matter to ask several questions so that the children don't have to read too much for each question. For instance the example for testing the understanding of concepts in Resource 15 could be the basis for a process skill question such as:

"If the snails had not moved in straight lines, how would you have measured the distance they had moved?"

This would be a useful question in a test where both skills and concepts were being tested because answering this does not depend on answering the other questions in the item correctly. It should be marked separately, however, and not as part of the concept-based question.

Task F5

Organisation: Participants in pairs.

Materials needed: Task F5, Resource 15, locally used curriculum or syllabus.

Time: This activity can take as long as is available. If time is limited allow about 60 minutes workshop time and ask for other items to be created outside the workshop.

Instructions: Tell participants to:

- Draft items and mark schemes which could be used in schools in their locality for pupils in the fifth or sixth years of primary school.

Task F5
Select a part of the current science curriculum for pupils in the fifth or sixth year of primary school and prepare items to assess understanding of the relevant ideas and the skills that are intended to be developed. Prepare mark schemes for these items in a suitable form.

Plenary discussion

This can take the same form as for Task F4. It may be useful to add to the earlier point about using the same context/subject matter for several questions. When a test is being created, it is often possible to use the same context (situation) and to ask separate questions requiring process skills to be used. (See the example at the end of background information) Although this reduces the amount of reading it increases any bias that may be associated with this context.

The development of questions is not complete until they have been tried out with children. If possible arrange for this as part of the workshop, using a nearby school. If not, ask participants to try them with children between sessions of the workshop and bring the responses to the next session.

Points to emphasise

- It is the mark scheme which will ultimately determine what is given credit. Thus it is very important to develop this at the same time as the item. In the case of open-ended questions it is helpful to do this in two phases. First produce a draft mark scheme, then try out the items with children and adjust the scheme as necessary. Often children produce types of answers which are not at first anticipated and may make a difference to what is considered worthy of marks.
- The question should give some indication to the children as to what is expected as an answer. For example if 'similarities and differences' are asked for, then specify the number expected and create the mark scheme to fit. Again, in a planning question, if the mark scheme gives credit for mentioning, say, equipment to be used and how the results will be calculated, then make this clear in the question: e.g. 'Make sure you say: what you will use, how you will use it, and how you will decide the result...'

Background information

There are two main ways of arriving at a summary assessment of children's achievement at a particular time. One is by forming a summary of the records of on-going assessment; the other is by 'checking up' through giving some special tasks, or test items, which are designed to find out the extent to which children have developed certain skills and ideas. Ideally both would be used, so that the special tasks become checks for the teachers' assessment based on records of formative assessment.

As discussed in Module E, summarising can preserve the detail of the on-going assessment and enable summary statements to be made about progress during the period in question over a wide range of ideas, skills and attitudes. At the same time two less favourable aspects need to be considered: first, that some of the information may be out of date if areas of work at the beginning of the year have not been revisited later; second, that it may be difficult to produce a descriptive summary which enables pupils' achievement to be compared, for their opportunities may have varied.

Checking up has the attraction of providing apparently equal opportunities for children to show what they can do at a particular time, although necessarily over a reduced range of experiences. However, it must be remembered that giving children the same tasks is not the same as giving them the same opportunities since learning opportunity implies a match between the task and the child's ability to engage with it. In other words, children who have developed the ability being tested may not be able to show this in the test situation.

Issues relating to tests

Here we consider the approach to summative assessment of giving special tasks or tests. In doing so their disadvantages should be kept in mind:

- They have low validity since the tasks are restricted to those which can be done in the test context.
- There is a restricted range of skills and ideas that can be included, partly because of the limitation on the length of a test and partly because some skills are only revealed in extended activities.
- They require considerable teacher time for organisation and marking.
- They take up learning time.
- They can induce anxiety in children.

To minimise the disadvantage in terms of validity the test items must cover the full range of skills and ideas that are the goals of science education, but in less detail than in the case of formative assessment. This means that it must go further than testing a knowledge of facts, which is always relatively easy. For example:

What is the boiling point of water?

- | | |
|---|-------|
| A | 0°C |
| B | 32°C |
| C | 90°C |
| D | 100°C |

What is the name of the part in the centre of the eye through which light passes?

Would these items really tell us anything about children's scientific development? The knowledge of facts is a means of building broader generalisations, the ideas or concepts which constitute the more important knowledge that children need to make sense of their environment. Testing knowledge of facts must not replace testing the understanding of

ideas; this shows in the ability to apply the ideas in situations other than those in which they were learned.

For checking on children's developing ideas the important aim is to ensure that children are applying their ideas, working out their answer from their own thinking rather than simply recalling it. The decision as to what children are likely to be able to recall and what they will have to work out depends on the experiences of the children. For some, an answer to the question: *How would you try to stop an ice cube from melting for as long as possible without using a 'fridge?* may require application of ideas about the need for heat to bring about melting and transfer of heat through different materials. For others, however, who may have undertaken such an activity, it may be a matter of recalling what was done and the reasons put forward for it at that time. However there are many variations in ways of ensuring that the application of ideas is required and it should be possible to avoid questions which ask for straight recall of facts or for applications which can be recalled.

Forms of presentation of test items

Test questions are most commonly presented on paper, perhaps using pictures or symbols, and pupils answer by writing or drawing. This is more suitable for assessing ideas or concepts than for process skills. To assess practical enquiry skills it is necessary for pupils to be given tasks to be carried out using real objects and equipment. This is particularly so for assessing the performance of whole investigations.

However, the process of practical enquiry can be divided into component skills, such as planning procedures, identifying variables, making observations, recording findings and drawing conclusions from them. Some of these components lend themselves to assessment on paper and the difficulty of providing practical test situations means that this is an attractive option. Two points must be kept in mind when assessing process skills on paper:

- for a skill to be assessed the pupils must be required to use the skill, not just to show knowledge of how to use it
- being able to perform separate components of investigations in isolation from the whole is not the same as performing a whole investigation from start to finish.

Forms of answering test items

In all assessment the influence on performance of abilities which are not under test should be minimised. So, particularly for young children, it is important that the demand of a question in terms of reading and writing should not prevent them showing what they understand or showing the scientific skills they have. For young children it is helpful for them to be able to respond by drawing, constructing something or carrying out an investigation.

In written questions, one of the popular ways of reducing the burden of writing – and at the same time making marking easy is to provide alternative answers from which the pupil chooses an answer. These multiple-choice tests have many disadvantages, however, in relation to assessing learning in science.

A multiple-choice question takes the form of a 'stem' which is a statement or question and several alternative answers. For example :

Tigers are striped so that they can: A keep clean B run faster C hide easily from other animals D be seen easily by other animals

Some of the points in favour of such questions are:

- they do not require pupils to write answers and so do not limit pupils' performance on account of writing ability
- they can be marked easily, by hand or by machine
- they can be answered quickly so that several such questions can be asked and a range of different topics covered.

Points against questions in this form include:

- they can often be answered by intelligent guessing or elimination of obviously wrong alternatives (distracters)
- they do not assess what pupils can produce as answers, only their recognition of right answers (plus the ability to eliminate wrong distracters and to 'spot' the signs of correct answers)
- they give no guarantee that the ideas behind the questions have been understood; the right choice might be made for the wrong reason
- they often require more reading than open-ended questions, which is time-consuming and causes difficulty for some pupils
- they are difficult to produce and need careful trials to eliminate ambiguous wording and supposedly wrong answers which turn out to be correct, or vice versa
- young children find this form of question difficult to handle; they often tick more than one answer, however careful the instructions
- they do not test the ability of learners to express themselves, which is an important aspect of science education.

Marking

One of the chief reasons for the introduction of multiple-choice questions was to avoid difficulties in marking. However, all that this did was to shift the balance of difficulty to the writing of the questions. Open questions are much easier to write but more difficult to mark. For example the question above could be expressed in a more straightforward way:

What advantages do tigers get from being striped?

.....

But this would be more difficult to mark. Inevitably pupils will express their ideas in different ways and most answers will be neither wholly correct nor wholly incorrect. There are well established ways of dealing with such answers, which involve establishing qualitative categories of answers. For example, such categories for the open question about the tiger's stripes might be:

- An answer which gives correct reasons making reference to
 - stripes not easily seen in tigers' habitat
 - therefore they can hunt without being seen or attacked
 - or • other ways of expressing the idea of camouflage
- An answer which is correct as far as it goes, but is limited eg the single word 'camouflage'
- Incorrect answer, eg reference to stripes frightening other animals
- Not answering the question.

The boundaries of the categories would be derived either from the criteria to be used, if these exist, or from the kinds of answers which can be expected of the children of the age being tested. When the general compass of the categories has been decided it is useful to

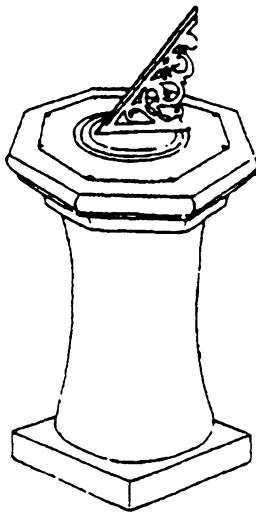
provide examples of answers which would fall into them. So, for instance, a decision would have to be made as to where an answer such as '*it helps them catch their food*' would fall. If illustrative responses as well as descriptions for each category are provided then it is entirely feasible for marking schemes of this type to be used reliably.

For some purposes, it is most useful to report in terms of the qualitative categories. This gives information about the extent to which pupils have achieved certain ideas. For other purposes, marks could be assigned to the categories (eg a = 2, b = 1, c and d = 0).

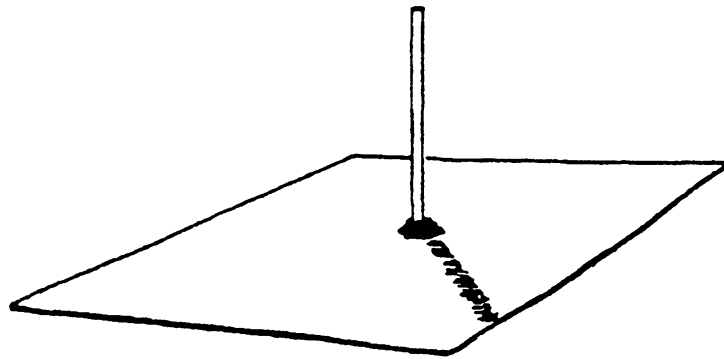
Example of a set of questions on the same theme

(Taken from Hargreaves, et al 1990)

After looking at a sun-dial in an ornamental garden, children made their own:

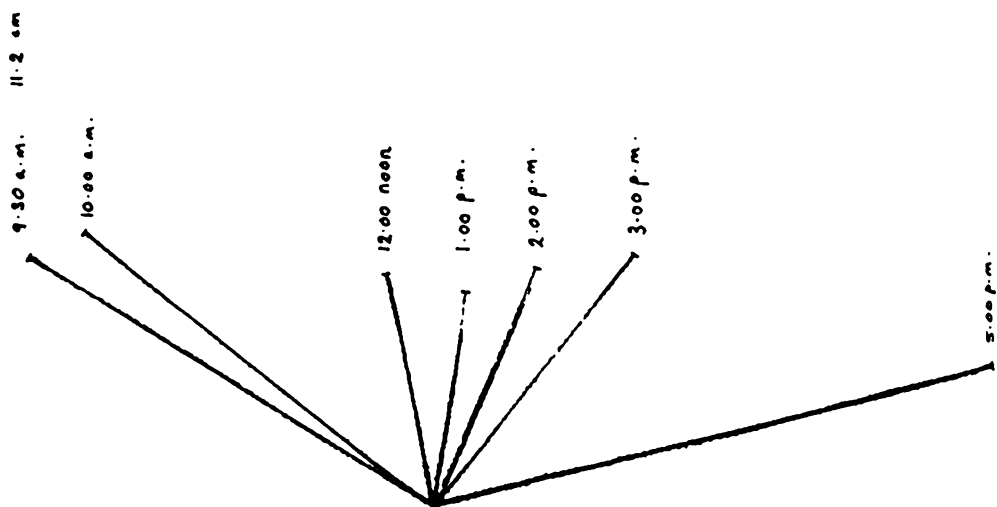


The sun-dial in the garden



The children's sun-dial

The shadows on this page were marked by children during one day in August. They used the sun-dial they had made.



1. Measure the shadow length at each time shown. Write the time and length in the table below.

The first one has been done for you.

Time of day Length of shadow in cm

9.30 a.m.	11.2

2. At 4.00 p.m. it was cloudy, so there was no shadow.
Draw in the shadow for 4.00 p.m. as you think it might have looked.
3. How did you decide where and how to draw the shadow?