Assessment: principles and approaches

WORKSHOP LEADER'S NOTES

Aims

This is an introductory module which aims to introduce participants to the whole course and then to share preconceptions about assessment as a basis for developing a wider understanding of the nature and roles of assessment.

Objectives

1. Understanding of the meaning of assessment in various learning contexts.
2. Knowledge of the various purposes of assessment.
3. Awareness of the range of information and ways of collecting it that are useful for assessment in various contexts.
4. Understanding of the concepts of reliability and validity.
5. Recognition of various issues, such as gender bias, in assessment.

Introduction to the course

- Give the participants an overview of the topics that will be covered in the modules selected for their course. Use the notes on page 2 for this purpose.
- Tell them that the way they will work is through workshop activities, which means that they will be undertaking tasks, mostly in groups, and discussing ideas about assessment through studying examples. There will be no lecturing; they will have access to background information at various points which they may read between sessions.
- As part of each workshop task there is some kind of reporting to others, and groups will be asked to bring their ideas together in a form that can be shared with others.

Note: The background information is for the workshop leader to read thoroughly before the module and to use in the discussion. Do not copy it for participants until they have completed the complete module. Then it may be useful for them to read sections or the whole to consolidate the points that have been made.

Introduction to Module A

- Indicate to participants the aims of the module, emphasising that it is introductory and that many of the aspects of the subject introduced can be studied in more depth in other modules later.
- Active involvement of the participants is important from the start, so keep the introductory remarks brief.

Task A1

This task provides an opportunity for participants to discuss basic points about the meaning and methods of assessment. It is important for every participant to consider his/her own ideas and to listen to others' ideas.

Organisation: Participants in groups of about five. Whilst they are working, walk round the groups and check for any misunderstanding of the task (for example if there is a query about kinds of information, give an example such as about the children's observation, use of measurements, co-operation, etc.). Make sure that there is discussion in the groups and that someone is making a record of points to report.

Materials needed: Task A1, Resource No 1.

Time: Allow at least 30 minutes for the task and 30 minutes for plenary feedback.

Instructions: Tell the participants to:

- Read the resource and the task.
- Make a list of aspects of the children's scientific development the teacher could gain information about in this sequence of events.
- For each one, note how the information would have been gained by the teacher.
- Ask themselves whether in their view these aspects were being assessed by the teacher.
- What else would the teacher need to do to assess these aspects?

Task A1

Read the account in Resource 1 of how an investigation about heartbeat rates was planned and undertaken by pupils in the fifth year of primary school. Then, working as a group:

(i) List all the kinds of information relevant to the children's scientific development that the teacher gained in the events described. (It may help to focus on one child, the one called J.)
(ii) Indicate how you think these kinds of information were gathered by the teacher.
(iii) What aspects of development were assessed in these activities?
(iv) What else would the teacher need to do to assess the children's achievements?

In each case note points about which there is disagreement as well as agreement within the groups. Prepare to report on your answers and discussion to other groups.

Plenary discussion

Ask one group's spokesperson to report on sub-task (i). Then ask other groups to add or to challenge what this group has reported, but not to report in full.

Discuss the list in terms of the evidence for the information which is claimed. If you disagree with an item ask, 'which particular event would give that kind of information?'

Then ask another group to report their response to sub-task (ii) and collect additional or differing responses from other groups.

Repeat this procedure for sub-tasks (iii) and (iv) introducing some of the following points at appropriate times.

Points to emphasise

Opportunities to underline these points are likely to arise during the feedback and discussion, but if not the leader should introduce them.

- Assessment is a broad term which is used to cover judgements about performance made both in informal classroom contexts and in more formal, test situations. (See background information for meanings of assessment).
- Some participants may only recognise as 'assessment' the more formal setting of questions at the end of the investigation. Ask them why. Is it because these concern the knowledge of the children rather than their investigation skills? Is it because asking for written answers is more formal than observing and listening to children? What other reasons?
- Gathering information, of itself, is not enough for assessment. Assessment is not just a description of what children can do but implies some judgement of it. All kinds of information are helpful for assessment, but assessment is more than information.

- Assessment means making a judgement as in: 'how well is this child doing in this aspect of development?' To make the judgement, the information has to be compared with something – 'how well . . . .' in relation to what? It could be:
  - how well compared with a specified type of performance (this is called criterion-referenced assessment) or
  - 'how well' compared with what that particular pupil could do at an earlier time (this is called pupil-referenced assessment) or
  - 'how well' compared with what is normal for pupils of the same age or stage to be able to do (this is called norm-referenced assessment).
- To summarise: all assessments involve some basis for judgement, thus:

  Information + basis for judgement —> assessment
- Use examples from the background information to underline the importance of using an appropriate basis for judgement. However, the next task takes up this matter in a specific context.

Task A2

This task provides an opportunity for further reflection on the meaning and of assessment and on the bases for judgement which might be used in particular instances. The task is open and there is no 'correct' answer since much will depend on the context and on the purpose for which the teacher was carrying out the assessment.

Organisation: Participants in groups of about five.

Materials needed: Task A2; Resource 2.

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

Instructions: Tell the participants to:
- List the information as quickly as possible (this is similar to what was done in Task A1 and is just a necessary step for getting to sub-task (ii).
- Consider the pros and cons or each basis for judging the information that is gathered. (Ensure that the meaning for pupil-referenced, norm-referenced and criterion-referenced are understood.)

Task A2

Resource 2 is an account of an activity deliberately devised to provide opportunities for assessing pupils' work in science. Put yourself in the position of the teacher and, with your group, answer these questions:

(i) What kinds of information would the teacher be able to collect about the pupils?

(ii) What basis for assessing their performance would be the most appropriate? Consider each of these:

Pupil-referenced – comparing the information with what that particular pupil could do at an earlier time

Criterion-referenced – comparing the information with a specified standard of performance

Norm-referenced – comparing with what is the norm (or average) for pupils of the same age or stage.

Plenary discussion

For sub-task (i) collect responses by writing them on a board or large sheet of paper, taking one item from each group in turn and asking them not to repeat what others have said, until all the responses have been collected.

Sub-task (ii), collect 'votes' for each of the three possibilities – pupil-referenced, criterion-reference, norm-referenced. Ask those choosing each one to justify their choice.

Points to emphasise

- Clearly, the teacher who devised the task intended to assess things which it was not easy to assess in other ways, as well as to make the whole process fun for the children. Thus it is important to include items such as *questioning, curiosity, ability to communicate information orally*, in the list.

- There is no 'right' answer to (ii), since the basis for judgement will depend on what *purpose* the teacher had in mind for the assessment:
  - If the teacher was looking for progress made by each pupil, then pupil-referencing would have been appropriate.
  - If the teacher wanted to see what the children could do, using a standard which was the same for all the children, then it would be preferable for this to be based on the criteria of performance, ie the assessment would be criterion-referenced; it could hardly be based on a norm.
    - There is no place for norm-referenced assessment in teachers' assessment of the kind described in this example. Norm-referencing can only be used where a large number of pupils have been given the same task so that an average performance has been identified.

- The discussion of using criteria raises the question of 'which criteria?' Module B aims to introduce and give practice in using some criteria, so discussion of this point should be delayed. For the moment there is the important matter of the reliability and validity to consider.

Task A3

This task is a quick one to do, but leads to discussion of important matters relating to reliability, validity and gender bias.

Organisation: Participants in groups of about five.

Materials needed: Task A3, Resource 3.

Time: Allow 10 minutes for individual work, 15 minutes for sharing in the groups and 20 minutes for the plenary discussion.

Instructions: Tell the participants to:
- Answer the questions (i) to (v) individually and quickly.
- Share answers in the group and note points of agreement and disagreement, reasons.
- Be prepared to report the group's answers.

Task A3

Read the three tasks, A, B, and C in Resource 3 which assess in different ways children's ability to investigate.
Then answer these questions and give a reason for your answer in each case.
Do this individually and then share your views with the group.

(i) Which task (A, B or C) do you think is the closest to assessing how well the children can investigate something?
(ii) Which one gives a result that would be most likely to be judged in the same way by different teachers?
(iii) Which is the most interesting from the point of view of the children?
(iv) Which is the most convenient for the teacher?
(v) Which task would be best for finding out what the children would do to investigate, even if the teacher can't let them do it?

Don't forget to give the reason for your choice!

Plenary discussion

Ask one group for its response to question (i). Note any disagreement from other groups. Ask another group for its response to question (ii) and so on. Introduce the following points during the discussion.

Points to emphasise

- In relation to the reasons given in answer to (i), introduce the idea of *validity*. This word is used to describe the extent to which what the children were asked to do, really gives them the opportunity to show the skills or idea, or whatever was being assessed (see background information). So, Task A asks the children to *recognise and select* what would be the best test for bendiness, but they do not produce the ideas themselves. Task B asks them to plan, but not carry out the investigation, so it assesses planning only rather than ability to investigate. Task C enables the children to show how they would investigate and so would be the most valid assessment in this case.
- Discussion of the answers to question (ii) provides an opportunity to introduce the word and concept of *reliability*. Reliability relates to the accuracy of the assessment and means how closely the result of the assessment would be the same if the assessment were repeated. In the case of the assessment tasks in Resource No 3 it means the extent to which different teachers would make the same judgement of the pupils' answers. Since the 'correctness' of the answer can be unambiguously judged in Task A this has the highest reliability. Task B might be next, since all the information is written down. Assessment of Task C would depend on observing actions as well as reading what has been written.
- Note that reliability and validity *don't* go together. There usually has to be a trade-off of one against the other. What this is will depend on the purpose of the assessment. For assessment which is intended to be diagnostic and help teaching and learning, validity is paramount and reliability of less importance. But if the assessment is to be used for selection or comparison of children, then reliability is important and has to be increased as far as possible without infringing validity too much. (There is more about this in the background information for Module E).
- In discussing answers to question (iii), note that interest is not a trivial matter. If a task is enjoyable to the children this means that their attention and effort are likely to be engaged and so this adds to validity. When the attractiveness is different for groups of children, such as boys and girls, the result may be biased in favour of one group. The subject matter of the task is thus important, even if direct knowledge of it is not required. For example, if the activity about camouflage in Resource 2 had been about camouflage

of soldiers in war rather than about animals it might be less interesting to girls than to boys (see background information).

- In the discussion of question (iv) note that good multiple-choice questions are not easy to write, although they may seem most convenient to use. Example A is not a good question since there is more than one right answer. (The pros and cons of multiple choice questions are taken up in background information for Module F).
- In discussing question (v) note that it is not always possible to use practical items, even if this is preferred and thought to be the most valid. The important point is not to interpret the result as if it were the result of practical activity when it is not. Points from the section of background information on *interpreting results of assessment* can be brought in here.

Background information

Meanings of assessment

It is generally agreed that assessment in the context of children's achievements in school is a process of making judgements about the extent of these achievements. The judgements are reached on the basis of information which has been gathered about performance and which is compared with some kind of expectation. The various ways in which information is collected and the various bases for judging it, create the variety of different kinds of assessment. These include, at one extreme, standardised tests, where information is gathered whilst children are tackling carefully devised tests under controlled conditions and, in contrast, ongoing assessment, carried out almost imperceptibly during normal interchange between teacher and pupils.

The major distinction within assessment methods is between tests (and examinations) and other forms of assessment. Indeed some use of the term 'assessment' excludes tests and means only various forms of informal assessment usually devised by, and always conducted by, the teacher. Tests are specially devised activities designed to assess knowledge and/or skills by giving precisely the same task to pupils, who have to respond to it under similar conditions as envisaged by those who devised and trialled the test. However the distinction between tests and non-test assessment is not always not very clear. Some 'tests' can be absorbed into classroom work and look very much like normal classroom work as far as the children are concerned and so they cannot always be regarded as formal. To be more useful, the distinction should go beyond methods to include purposes. The main purpose of tests is to check up what children have achieved, although in some cases they also serve a purpose of feedback to help learning.

Evaluation and assessment

The terms 'evaluation' and 'assessment' are used differently in different countries. In some cases both are used interchangeably in relation to pupils' achievements. However, in other cases the word 'assessment' is used in the context of making judgements about pupils' achievements, and 'evaluation' when making judgements about other things, such as the curriculum, teaching materials and methods. In these materials we follow the latter convention and talk about assessment of pupils, rather than evaluation.

Purposes of assessment

The range of purposes of assessment of can be organised under headings such as the following:

- formative, or ongoing, to help teaching and learning
- summative or summary, to indicate achievment at a certain point
- selection or certification
- school evaluation
- national monitoring

Assessment in each of these categories requires information which fits its purpose. The full list is given for completeness, although our main concern is only with the first two.

Formative assessment is aimed at helping the teaching and learning process; information gathered regularly is used for making decisions during ongoing work. It assists teachers in adjusting the challenges given to children to match their existing ideas and skills, to help rather than to grade children. It is usually informal in that the child is not aware that it is taking place.

The term **diagnostic assessment** is sometimes used as if this were different from formative assessment. Diagnostic assessment has a more specific focus, being concerned with

examining in depth a particular area of performance. But this is only a slight variant of formative assessment and can be considered part of it.

Summative assessment, as the name suggests, means a summary judgement or a summing up of where a child has reached at a certain time. Quite often the information is obtained by a test (or examination) at the end of a term, year, or of a certain section of work. But it is also possible to give a summative assessment as a result of reviewing records of ongoing assessment, as teachers frequently do in reporting to parents, either orally or in writing, at the end of a year.

Methods

Ways of collecting information about children's achievement of ideas, skills and attitudes can be categorised in terms of what the children are doing when the information is collected and how the information is collected.

The children may be engaged in:
- normal work (including both written and practical work)
- special practical tasks (including tests)
- special written tasks (including tests)
- self-assessment

and the teacher may be:
- observing, but not interacting with, the children (including watching and listening)
- interacting with children (as well as watching and listening)
- using a check-list
- marking tests
- reading or marking class work
- gathering general impressions.

Common combinations of items from these lists describe identifiable 'methods' of assessment such as tests, continuous assessment and ratings, but there is clearly a range of other possibilities. Some methods are more suited than others to collecting information about achievement in particular subject areas and so here we focus on methods particularly appropriate for performance in science and technology.

Basis of judgement

This refers to the reference point used in judging information. It may be illustrated in terms of an example (from Harlen, 1996):

> Suppose, as a hypothetical example that a teacher wants to assess a child's ability in 'knocking nails into wood'.
>
> The teacher may have some expectation of the level of performance (knocking the nail in straight, using the hammer correctly, taking necessary safety precautions) and judge the child's performance in relation to these. The judgement is made in terms of the extent to which the child's performance meets the criteria; that is, it is *criterion-referenced*.
>
> Alternatively the teacher may judge in terms of how the child performs at knocking in nails compared with other children of the same age and stage. If this is the case there will be a norm or average performance known for the age/stage group and any child can be described in relation to this as average, above average or below average, or more precisely identified if some quantitative measure has been obtained. (The result could be expressed as a 'knocking nails age' or a 'hammer manipulation' quotient!) The judgement arrived at in this way is called a *norm-referenced* assessment.

A third possibility is that the teacher compares the child's present performance with what the same child could do on a previous occasion – in which case the assessment is *pupil-referenced* (or ipsative).

It is important to recognise these different bases for judgements in assessment and apply them appropriately. They each have their value in the right context, but each have drawbacks outside these contexts. Pupil-referenced assessment is appropriate for formative assessment, for providing encouraging feedback to pupils, particularly slower ones who, if compared with criteria or with others' performance would always be seeming to fail, but can recognise progress in terms of their own previous performance. But it must be realised that it leads to one child being praised for work which, from another child, might be received with less approval. This is no problem as long as no comparisons are made between children, but where comparisons are being made, or performance in terms of external standards has to be reported, then one of the other bases for judgements must be used.

Interpreting results of assessment

Results of assessment have to be interpreted in the knowledge of the kind of information gathered and the basis of the judgement made. There is always some implied generalisation of the result, for we assess only a sample of behaviour and act on it as if it applied to more than this sample. Indeed unless we are assessing simple recall of facts, we have the expectation that the result will tell us more than just about the particular performance assessed. However, it is important not to generalise beyond what is justified by the information. But how is this limit to be identified? The problem can be put in this way: if a child has shown evidence, either in a test situation or in regular work, of using patterns in findings to make predictions, to what extent does this mean that (s)he is able to do this on all other occasions where it is possible?

It is known from research evidence that context influences performance, but to an unknown extent. Assessment results must therefore be interpreted cautiously, as guides to what children can do, but not as indicating any kind of certainty about it. No assessment can be used predictively with certainty; it is best treated as a hypothesis, a tentative finding, to be modified by further evidence of the child's performance.

Cautious interpretation should avoid labelling children, which results from over-generalisation of assessment results. When an assessment is interpreted as if it describes the whole child and not just certain performance this can affect teachers', parents' and the child's view of what (s)he is able to do, often needlessly limiting expectations.

Bias relating to gender

An important consequence of the influence on performance of the context and the subject matter is that it can disadvantage certain groups of children. Some children may be less motivated than others by the subject they are asked to think about. Some may anticipate failure in certain activities because of their self-image and so make failure more likely. The poor performance of girls, in some activities involving ideas of physics, can be explained to an extent by the context and topics relating more to boys' interests and triggering the reaction in girls 'I can't do this' before they even try.

There is also evidence that, in tests and examinations, boys perform better on multiple-choice questions than on ones where an open response must be written and that for girls the reverse is the case. Girls have also been rated as less willing to undertake investigations and so may achieve less well in a practical context than boys, particularly if they have, in class, been used to taking a passive role in science and technology activities.

Reliability and validity

There are two concepts in assessment which describe how dependable an assessment is. These are its reliability and its validity.

Reliability is the term used to express the degree of accuracy of the result of assessment ie, if the assessment were to be repeated, the extent to which the second result would agree with the first. The reliability of assessment of a child's skill in writing is likely to be less than the reliability of a test of addition and subtraction sums, mainly because it is far easier to mark the latter consistently.

Validity refers to the extent to which what is assessed really reflects the behaviour it was intended to assess. For example, a multiple-choice test of knowledge about materials that conduct electricity would not give a valid assessment of understanding of a simple electric circuit, although it would be an assessment of quite high reliability.

The highest possible level of reliability and validity is always the aim, but since neither can be 100% and since the requirements of one sometimes conflict with those of the other, a compromise has to be accepted. The requirements in terms of relative emphasis on detail, and on reliability and validity will vary with the purpose of the assessment. For example, for formative, ongoing, assessment, it is more important that the result has a high validity, indicating what the child is and is not yet able to do. For the purposes of selection, or wherever children are compared one with another, it is important that the assessment is reliable, whilst at the same time being valid in terms of the abilities being assessed.