Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: https://www.tandfonline.com/loi/caie20

Purposes and Procedures for Assessing Science Process Skills

WYNNE HARLEN

To cite this article: WYNNE HARLEN (1999) Purposes and Procedures for Assessing Science Process Skills, Assessment in Education: Principles, Policy & Practice, 6:1, 129-144, DOI: 10.1080/09695949993044

To link to this article: https://doi.org/10.1080/09695949993044

Published online: 09 Jun 2010.

Purposes and Procedures for Assessing Science Process Skills

WYNNE HARLEN
Scottish Council for Research in Education, 15 St John Street, Edinburgh EH8 8JR, UK

ABSTRACT *Science process skills are inseparable in practice from the conceptual understanding that is involved in learning and applying science. Nevertheless, it is useful to identify and discuss the skills which can apply to different subject-matter because of their central role in learning with understanding, whether in formal education or throughout life. That role is also the reason for the importance of assessing the development of science process skills. This paper argues that it is a content-dominated view of science education rather than the technical difficulties that has inhibited the development of effective procedures for assessing process skills to date. The discussion focuses on approaches to process skill assessment for three purposes, formative, summative and national and international monitoring. In all cases the assessment of skills is influenced not only by the ability to use the skill but also by knowledge of and familiarity with the subject-matter with which the skills are used. Thus what is assessed in any particular situation is a combination of skills and knowledge and various steps have to be taken if these are to be separated.*

Introduction

The subject of this paper is the assessment of those mental and physical skills variously described as scientific process skills (Organisation for Economic Co-operation and Development (OECD), 1998), procedural skills (e.g. Gott & Duggan, 1994), experimental and investigative science (National Curriculum; Department for Education, 1995), habits of mind (American Association for the Advancement of Science, 1993) or scientific inquiry abilities (National Academy of Sciences, 1994). Despite differences in the generic title, there is considerable agreement about what these include at the more detailed level. Some subsume values and attitudes but all include, in one form or another, abilities relating to identifying investigable questions, designing investigations, obtaining evidence, interpreting evidence in terms of the question addressed in the inquiry, and communicating the investigation process. There have been arguments about whether it is sensible to identify these abilities as separate skills (e.g. Wellington, 1989) but it is widely understood that they are important aspects of science education and that by discussing them separately as such, no claim is being made that they can be isolated from the content or subject-matter of science. Indeed, science process skills are best thought of as a face

0969-594X/99/010129-16

of a solid three-dimensional object, integral to the whole and a real and tangible part of the whole but having no independent existence, just as the surface of a solid has no reality except as part of the whole and yet can be described, changed and evaluated.

The Importance of Science Process Skills

All skills have to be used in some context and scientific process skills are *only* scientific if they are applied in the context of science. Otherwise they are general descriptions of logical and rational thinking which are used in many areas of human endeavour. Not only are science process skills always exercised in relation to some science content, but they can relate to the full range of science content and have a central role in learning with understanding about this content. It is precisely for this reason that it is important to consider them separately.

Learning with understanding involves linking new experiences to previous ones and extending ideas and concepts to include a progressively wider range of related phenomena. In this way the ideas developed in relation to particular phenomena ('small' ideas) become linked together to form ones that apply to a wider range of phenomena and so have more explanatory power ('big' ideas). Learning with understanding in science involves testing the usefulness of possible explanatory ideas by using them to make predictions or to pose questions, collecting evidence to test the prediction or answer the questions and interpreting the result; in other words, using the science process skills. Conceptual learning conceived in this way takes place through the gradual development of ideas that make sense of the experience and evidence available to the learner at a particular time. As experience is extended and current understanding is no longer adequate, the process continues and alternative ideas, offered by the teacher or other pupils or obtained from other sources, may be tested. The role of the process skills in this development of understanding is crucial. If these skills are not well developed and, for example, relevant evidence is not collected, or conclusions are based selectively on those findings which confirm initial preconceptions and ignore contrary evidence, then the emerging concepts will not help understanding of the world around. Thus the development of scientific process skills has to be a major goal of science education. This is now acknowledged in curriculum statements world-wide.

It is tempting to conclude that this is good enough reason for including science process skills in any assessment of learning in science. If this were the case then there would be many more examples of process skills in national tests, examinations and teachers' assessments than in fact are found in practice. Part of the reason for this is the technical difficulty of assessing some of the process skills, but the main reason must surely be the inhibiting influence of a view of science education as being concerned only with the development of scientific concepts and knowledge (Tobin *et al.*, 1990). The technical problems can be solved where there is a will to do so. But first it is necessary to counter the argument that science education is ultimately about understanding, that using science process skills is only a means to that end and thus only the end product needs to be assessed. This denies the value of these skills in their own right and ignores the strong case for process skills being included as major aims of science education. This case is made, not just in terms of preparing future scientists, who will be 'doing' science, but in terms of the whole population, who need 'scientific literacy' in order to live in a world where science impinges on most aspects of personal, social and global life.

Scientific literacy has been defined in many different ways. The most recent is the outcome of the thinking of the 'science functional expert group' set up to design the framework for the OECD surveys of student achievement (the Performance Indicators in Student Achievement (PISA) project). In this project the focus is on the outcomes of the whole of school education in the compulsory years and tests of reading, mathematics and science are planned for the years 2000, 2003 and 2006. The surveys of students in their last year of compulsory education (15-year-olds) are designed to answer the question of how education in each country is providing all of its citizens with what they need to function in their lives. In science this is identified as scientific literacy, defined as follows:

> By scientific literacy we mean being able to use scientific knowledge to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity. (OECD, 1999, in press)

The importance of developing thinking skills is also gaining support from research findings, particularly the Cognitive Acceleration through Science Education (CASE) project (Adey & Shayer, 1990, 1993, 1994; Shayer & Adey, 1993). There is also a rapidly growing emphasis world-wide on the development of 'core' skills (sometimes called 'key skills' or 'life skills'), which are seen as necessary to make lifelong learning a reality. Science has a key role to play in developing skills of communication, critical thinking and problem-solving and the ability to use and evaluate evidence. Thus assessment of the development and achievement of these important outcomes has to be included in the assessment of learning in science.

Challenges to Assessing Science Process Skills

As already mentioned, skills have to be used in relation to some content. Therein lie the difficulties in assessing these skills, for the performance in any task involving the skills will be influenced by the nature of the subject content as well as the ability to use the skill. A simple example illustrates the point. A 10-year-old student might well succeed in designing and carrying through an investigation of whether the height to which a ball rebounds after being dropped depends on the type of surface on which it is dropped; yet the same student is likely to be unable to succeed if the question to be investigated concerns the effect of concentration on the osmotic pressure of a certain solution. The Assessment of Performance Unit (APU), which surveyed national achievement in England, Wales and Northern Ireland from 1980 to 1985, using banks of items for assessing process skills to randomise the effect of the content, provides ample evidence of this effect (e.g. Department of Education and Science (DES) 1984a,b).

Moreover, the 'setting' or context of the task also influences performance, as it does in the assessment of the application of concepts, since a school or laboratory setting may signal that a particular kind of thinking is required whilst an everyday domestic setting would not provide this prompt. Again, APU results provided evidence of the influence of this aspect and of its unpredictability; in some cases an everyday context produced an easier task and in others a more difficult one (DES, 1988).

However, the extent to which the context, rather than cognitive demand, *dominates* the performance, as opposed to having a slightly modifying effect, is a matter of contention. It relates to the more general issues surrounding the conflicting claims made about the influence of context on learning, focusing on the notion of 'situated cognition'. Those who embrace this notion take the view that the context of a learning activity (not only its subject-matter but also the situation in which it is encountered) is so important that 'it completely over-rides any effect of either the logical structure of the task or the particular general ability of individuals' (Adey, 1997, p. 51). Evidence in support of this often cites examples such as the ability of unschooled young street vendors to calculate accurately whilst the same young people fail the arithmetic tasks if they are presented in the form of 'sums' at school. Similar findings have been reported for older people (Nunes *et al.*, 1993). Certainly the effect of context on performance has been demonstrated in many studies, notably by Watson & Johnson-Laird (1972). However Adey & Shayer (1994) and Adey (1997) have given spirited responses to these studies, including the well-known one of Donaldson (1978), who challenged Piaget's finding by showing that a change in context affected the difficulty of a Piagetian spatial ability task. Adey argues:

> Clearly motivation and interest play a significant role, and indeed if one wants to get a true picture of the maximum cognitive capability of a child it is essential that the task be made as relevant as possible, and that all of the essential pieces of concrete information are meaningful. But the determined efforts of the situated cognition adherents have failed so far to show that all conceptual difficulties can be accounted for simply by changing the context. (Adey, 1997, p. 59)

To extend Adey's argument, it may be that the variation associated with content or situation results from students' failure to engage with the cognitive demand. There is certainly evidence for this from talking with students about their perceptions of test items (e.g. Fensham, 1998; Fensham & Haslam, 1998). If the main problem is *engagement*, then the more this can be assured by providing tasks perceived as interesting and important to students, the less should be the variation associated with what the particular situations are. This is a challenge to 'authentic' assessment, which aims to provide assessment tasks that are 'more practical, realistic and challenging' (Torrance, 1995) than conventional tests.

However, the impact of conceptual understanding required when process skills are used in relation to scientific content remains. Clearly, it is not valid to assess process skills in tasks which require conceptual understanding not available to the student (as in the case of osmotic pressure and the primary student). It is important to assess process skills only in relation to content where the conceptual understanding will not be an obstacle to using process skills. This can never be certain, however, and the approaches to dealing with this problem effectively depend on whether the purpose of the assessment is formative or summative in relation to the individual or for monitoring standards at regional or national levels. These purposes will now be considered in turn.

Formative Assessment of Process Skills

Assessment is formative when information, gathered by teachers and by students assessing their own work, is used to adapt teaching and learning activities to meet the identified needs (Black & Wiliam, 1998a). Assessment which has a formative purpose is essentially in the hands of teachers and students and so, theoretically, can be part of every activity in which science process skills are used. Information can be gathered by:

- observing pupils—-this includes listening to how they describe their work and their reasoning;
- questioning, using open questions, phrased to invite pupils to explore their ideas and reasoning;
- setting tasks in a way which requires pupils to use certain skills; and
- asking pupils to communicate their thinking through drawings, artefacts, actions, role play and concept mapping, as well as writing.

These methods give plenty of opportunity for the teacher to find out the extent to which conceptual understanding is an obstacle to effective use of the process skills, or vice versa. Moreover, since the information is gathered in situations set up for students' learning, it is reasonable to expect the conceptual demand to be within the reach of the students; the teacher will certainly know if this is not the case. There is also the opportunity to gather information in a range of learning tasks and so build up a picture of development across these tasks. Hein (1991) and Hein & Price (1994) have pointed to a range of ideas for assessment, whilst stressing the importance of using a variety of approaches.

Teachers may, of course, collect information in the ways just suggested but yet not use it formatively. It is in the *use* of the information gathered that formative assessment is distinguished from assessment for other purposes. Use by the teacher involves decisions and action: decisions about the next steps in learning and action in helping pupils take these steps. But it is important to remember that it is the pupils who will take the next steps, and the more they are involved in the process, the greater will be their understanding of how to extend their learning.

To identify the focus for further learning, teachers need to have an understanding of development in process skills which they can both use themselves and share with their students. An example of how this development has been outlined for pupils aged 5–12 in a way which also suggests the next steps has been given by Harlen & Jelly (1997). For each process skill (observing, explaining, predicting, raising questions, planning and communicating) they suggest a list of questions reflecting as far as possible successive points in development. For example, for 'interpreting', the questions are:

Do the children

- discuss what they find in relation to their initial questions?
- compare their findings with their earlier predictions?
- notice associations between changes in one variable and another?
- identify patterns or trends in their results?
- check any patterns or trends against all the evidence?
- draw conclusions which summarise and are consistent with the evidence? and
- recognise that any conclusions may have to be changed in the light of new evidence? (Harlen & Jelly, 1997, p. 57)

Each list is intended to be used by the teacher to focus his or her attention on significant aspects of the students' actions. Reflecting after the event on the evidence gathered, the teacher can find where answers change from 'yes' to 'no' and so identify the region of development. Information used in this process does not only come from observing students; it can be provided by using all the methods listed above. Discussion with students of their methods of enquiry is particularly helpful. Written work is also a source of information if this is set so as to require reflection on methods. Peer review and groups reporting to each other enable teachers to hear students articulating their thinking.

Since assessment is not formative unless it is used to help learning, teachers and students have to be aware of ways in which this can be done. In the case of science process skills, the main strategies that *teachers* can use to help learning are:

- to provide an opportunity for using process skills; although obvious that this is necessary it is often not provided because students are given work cards or work books to follow and they do not have to think about what evidence is required and how best to gather it (justified because it enables students to cover the syllabus more quickly and is less demanding on the teacher);
- to encourage critical review by students of how their activities were carried out and discussion of how, if they were to repeat the activity, they could improve, for example, the design of the investigation, how evidence was gathered and how they used it to answer their initial question;
- to give feedback in a form that focuses on the quality of the work, not on the person (Black & Wiliam, 1998b);
- to give students access to examples of work which meet the criteria of quality and to point out the aspects which are significant in this (Wiliam, 1998);
- to engage in metacognitive discussion about procedures so that students see the relevance to other investigations of what they have learned about the way in which they conducted a particular investigation (Baird, 1998); and
- to teach the techniques and the language needed as skills advance (Roth, 1998).

The ways in which *students* can use this information are to engage in self-critical review of their work, take opportunities to compare it with examples of work of a higher quality and work to improve specific aspects. They should be helped to identify for themselves the steps they need to take and cross-check these with the teacher.

Since some process skills are, or should be, used in all science lessons—not only in laboratory exercises—there are frequent opportunities for assessment and for reflection on what is needed for further development. Moreover, it is not always necessary for the focus of formative assessment to be the individual student. When students work in groups and decisions about activities are made for the group as a whole, then the group is the focus of teaching and of formative assessment. Thus the burden on the teacher is not as great as the discussion above may seem to imply if it is assumed that the data gathering by observation has to be done for every student. However, where evidence about students, gathered for formative purposes, is later evaluated for summative purposes, then information about individual students is necessary. This 'double' use of the information is now discussed in the context of summative assessment.

Summative Assessment of Process Skills

In this section the concern is with assessment for the purposes of reporting on progress to date, or for certification, and involves comparing the performance of individuals with certain external standards or criteria. These may be standards set by the 'norm' or average performance of a group, or may be pre-determined as attainment targets at various levels.

One approach to arriving at a summative judgement is to use evidence already available, having been gathered and used for formative assessment. Converting this information into a summative judgement involves reviewing it against the standards or criteria that are used for reporting or deciding about levels of performance. To be specific, in the context of a curriculum with standards set out at levels (such as the National Curriculum in England and Wales), this means making a judgement about how the accumulated evidence, taken as a whole, matches one or other of the descriptions set out at the various levels (levels 1–8 in the National Curriculum). It is important to stress that it is the *evidence* that is used (i.e. the pieces of work which may be collected in a portfolio, or the notes of observations made), *not* records of judgements already related to levels. In other words, the process is one of review of *evidence* against criteria, not a simple arithmetic averaging of levels or scores. Levels add nothing to formative assessment where the purpose is to use the information to help teaching and learning, although the description of development that they embody, if they have been well constructed, may help in identifying the course of progression in skills and knowledge.

For summative assessment, however, agreed criteria in the form of standards or levels have an important role, first as a means of communicating what has been achieved, and second as providing some assurance that common standards have been applied. Fulfilling the two aspects of this role effectively requires that there is a common appreciation of what having achieved a certain level means in terms of performance and that similar judgements are made by different people about the level of performance shown across various pieces of work. A number of different ways of ensuring comparability in such judgements have been used and reviewed in Harlen (1994). Some form of moderation, where teachers can compare their judgements with those of others, is regarded as having considerable advantages, although exemplification is less costly and, according to Wiliam (1998), more effective in communicating operational meanings of criteria.

However, reaching a summative assessment using information collected for formative assessment has its drawbacks. It depends for its validity on opportunities having been created for students to show what they can do and on the teacher having collected relevant evidence. Unfortunately, it is all too common to find that students' use of science process skills is limited. There is massive evidence of 'recipe-following' science that gives no opportunity for thinking skills to be used or developed. Many primary teachers keep to 'safe' topics and keep children busy using work cards (Harlen *et al.*, 1995), whilst it has been reported that in secondary schools, laboratory exercises are often 'trivial' (Hofstein & Lunetta, 1982) or fail to engage students with the cognitive purpose of the activity (Marik *et al.*, 1990; Alton-Lee *et al.*, 1993). Indeed, there are many reasons why students may not be engaging in activities which give evidence of the science process skills and, even when they are, their teachers may not be able to assess them reliably.

Thus, it can be helpful to both students and teachers for special assessment tasks to be available to complement teachers' own assessments of process skills. This can help the students by giving them opportunities to show the skills that they have and it can help the teacher by giving concrete examples of the kinds of situations and questions that provided these opportunities and which could readily be incorporated in regular classroom activities. Practical experiences of this kind for primary pupils have been discussed by Russell & Harlen (1990), whilst a series of written tasks assessing science process skills is described in Schilling *et al.* (1990). At the secondary level, a major resource to help teachers in the assessment of practical work developed from a research project in Scotland on Techniques for the Assessment of Practical Skills. Three sets of materials have been produced, to assist with assessing basic skills in science (Bryce *et al.*, 1983), process skills (Bryce *et al.*, 1988) and practical investigations in biology, chemistry and physics (Bryce *et al.*, 1991). Banks of tasks have been developed by the Graded Assessment in Science Projects (Swain, 1991) and help for teachers has also been published by Gott *et al.* (1988), Woolnough (1991) and Jones *et al.* (1992).

Whilst individual practical skills can readily be assessed in a 'circus' style test, as, for example, in the APU tests (DES, 1982) and as described by Stark in the paper on the Assessment of Achievement Programme (AAP) in this issue, it has been argued that such an approach is 'educationally worthless (because it trivialises learning) and pedagogically dangerous (because it encourages bad teaching)' (Hodson, 1992). The emphasis, according to Hodson, should be on whole investigations. The APU tests did, and the AAP tests still do, include whole investigations and the results show conclusively that performance is strongly influenced by the subject-matter of the investigation. The bias according to subject-matter can be reduced if a sufficient range of different tasks involving different subject-matter can be used (Lock, 1990). When the tasks are practical ones this would mean a far longer test for each student than is likely to be acceptable. A test of reasonable length must necessarily cover fewer tasks and have a greater 'error' due to the particular choice of subject-matter. Thus for summative assessment of process skills to rely on such tests alone endangers the reliability of the assessment. The problem of length can be tackled by embedding the tasks in regular activities but this has drawbacks in terms of standardisation if important decisions for an individual rest on the results. The combination of a summary of on-going assessment and some well-designed practical tasks seems the best compromise for practical skills.

Using written tasks for assessing science process skills presents less of a problem in relation to context bias since more questions, covering a range of subject-matter, can be asked more quickly. The focus of concern is then validity rather than reliability. The arguments often focus around the use of multiple-choice items, favoured for summative assessment because of ease and supposed reliability of marking. But items in this form have been heavily criticised because of the ease with which even four or five distracters can be reduced, by the elimination of obviously incorrect statements, to a choice between two and thus a 50% chance of success by guessing. Further, correct answers are often chosen for the wrong reason (Tamir, 1990). However, even the reliability of such items is in doubt according to Black (1993), who claims that 'multiple choice questions are less reliable, for the same number of items, than open-ended versions of the same questions' (p. 71) and quotes Bridgeman (1992), who writes that, given the same testing time, open questions can produce as reliable a score as multiple-choice questions.

Many of the good examples of written questions which assess science process skills are derived from the pioneering work of the APU science project (Johnson, 1989). This is significant, since it is for national and international surveys that sufficient resources are made available for the necessary research and development to produce assessment matching the aims of modern science curricula. Current work of this kind, described in the next section, may well provide the stimulus that is needed to change assessment for individual students for summative and certification purposes.

It is important not to leave the subject of summative assessment without reference to its impact on the curriculum and on formative assessment practice. There is a strong tendency for changes in assessment to lag behind change in the curriculum and to act as a brake on innovation in the curriculum (Raizen *et al.*, 1989, 1990). When the assessment stakes are high for students and teachers, what is tested inevitably has a strong determining effect on what is taught. There is ample evidence that the introduction of summative assessment at ages 7, 11 and 14 in England and Wales has narrowed teaching (Pollard *et al.*, 1994; Daugherty, 1997), and the same influence is found in other countries (Crooks, 1988; Black & Wiliam, 1998b). The impact of testing has also influenced teachers' on-going assessment, turning this into a series of mini-summative assessments, where the emphasis is on deciding the level of achievement rather than using the information to assist further learning (Harlen & James, 1997). The consequence is that the benefits to be had from good formative assessment in terms of improving students' learning are lost. The situation can be rescued by developing wider understanding of the value of genuinely formative assessment whilst at the same time improving the match between curriculum aims and summative assessment procedures and instruments.

Assessing Process Skills in National Monitoring and International Comparisons

Monitoring at the national level involves the assessment of students as a sample of the population and has no immediate impact on their learning. Feedback is not to the individual student, teacher or school but into the system of education, through a long chain of events which involves politicians and advisers as well as those more directly concerned with education.

The purpose of monitoring at the national level is to find out the extent to which what it is intended that students should achieve is in fact being achieved across the country. The matter of what should be achieved is in many cases determined by the national or state curriculum, which may or may not include science process skills. The monitoring programme need not be confined to what is being taught, however, for it is appropriate for a monitoring programme to lead as well as to follow the curriculum, especially where the curriculum is under review.

In the case of the APU in England, Wales and Northern Ireland there was no national curriculum in place at the time of its inception in 1977 and so a consensus had to be created. This reflected the priorities of science education at the time, which gave a considerable emphasis to science process skills. Five of the six main categories of performance that were assessed related to process skills and three of these involved practical tests. The problems relating to the influence of subject-matter on process skill performance were overcome by having large banks of items for each skill and selecting randomly from these to produce many different test packages that were given to an independent random sample of pupils. The fact that any one student took only a small number of the full range of items testing a particular process skill was not a problem, since the results were of no consequence to the individual and were not reported at that level. The exception to the use of a large bank of items and reporting across tasks, however, was the assessment of performance on whole investigations. Despite intense effort with a variety of statistical techniques, it was not possible to form a useful 'score' across different investigations and so these were reported as separate tasks, describing in detail the range of achievement in various aspects of the investigations. For the other five categories of science, performance scores were produced which enabled year-on-year comparisons to be made and school variables to be explored in relation to various skills and sub-skills (Johnson, 1989). However, overall scores do not enable questions about the adequacy of performance to be answered. To find out what a score means in practice and consider how the measured performance might compare with expectations, it is necessary to look at individual test items.

The approach taken in the National Education Monitoring Programme in New Zealand is to report performance only for individual assessment items. No attempt is made to provide an overall score at the subject level or for groups of items relating to certain skills or concept areas. Science was one of the first subjects to be assessed in 1995 when this programme undertook its first survey. The results have been presented (Crooks & Flockton, 1996) in terms of percentages of different kinds of response for each reported task (not all tasks were reported, some being reserved for re-use in a subsequent assessment in a 4-year cycle). The performance of sub-groups formed by school and pupil background variables have been reported for each task. This approach has the considerable advantage that resources can be put into designing a small number of valid tasks, as opposed to creating a large bank. This has enabled the performance of individuals and collaborative groups of students to be video-taped and their performance analysed after the event. Other tasks have involved one-to-one interviews with students. The administration and marking calls heavily on experienced teachers and is valued as a professional development opportunity. The disadvantage of the approach is that judgement can only be made about performance task by task, with no overall picture which attempts to generalise across different topics and contexts.

An item by item approach would not serve the purposes of programmes of assessment set up to make international comparisons of achievement, such as those conducted by the International Association for the Evaluation of Educational Achievement (IEA) and the International Assessment of Educational Progress (IAEP) (see the papers in this volume by Johnson and Šetinc). The purpose of the IEA is to 'conduct co-operative international research studies in education' (Robitaille *et al.*, 1993) and the studies have provided information about the curriculum and teaching methods as well as performance in the participating countries. Nevertheless, the results that gain most attention are comparisons of achievement and how the test scores place countries in a league table of performance.

International surveys face a problem in arriving at a test comprising items that are regarded as fair by the participants. If fairness is taken to mean that students will have been taught what is tested then the test content has to be confined to the elements of the curriculum that are common to all countries. Inevitably, this results in innovative aspects, which may only appear in the curricula of some countries, being omitted. Consequently the IAEP study of 1991 and the first two IEA studies of science achievement, in 1970 and 1984, contained hardly any items assessing process skills. For the Third International Mathematics and Science Study (TIMSS), in 1995, the framework for assessing science identified three aspects to be assessed: content, performance expectations and perspectives. The performance expectations included science process skills as sub-divisions of five components: understanding; theorising, analysing and solving problems; using tools, routine procedures and science processes; investigating the natural world; and communicating. In the tests, however, in order to avoid going beyond the familiar acceptable to all countries, the vast majority of items assessed understanding.

The OECD PISA project, mentioned earlier (see p. 131), has taken a different approach to identifying what is to be tested. Rather than seeking consensus on what is taught in the science curriculum, it has sought agreement on what 15-year-old students should be able to do as a result of their science education. In setting up a framework for assessment the essential outcome of science education for all students has been identified as scientific literacy, defined earlier. Thus the results of the testing will indicate how well participating countries are preparing their students for the demands of their adult lives, rather than how well students have learned specific subject-matter. The PISA surveys should, therefore, be more forward-looking than the TIMSS surveys.

In order to assess the extent to which students have developed skills which can be used in everyday life, the PISA tests take the form of a number of tasks in which questions are asked about some stimulus material relating to real-life situations. The stimulus is authentic in that it concerns actual events in the present day or in history. The following example is part, only, of a task entitled ‘Stop that germ!’ The students are asked to read a short text about the history of immunisation. The extract of this text, on which two example items below are based, is as follows:

> As early as the 11th century, Chinese doctors were manipulating the immune system. By blowing pulverised scabs from a smallpox victim into their patients’ nostrils, they could often induce a mild case of the disease that prevented a more severe onslaught. In the 1700s, people rubbed their skins with dried scabs to protect themselves from the disease. These primitive practices were introduced into England and the American colonies. In 1771 and 1772, during a smallpox epidemic, a Boston doctor named Zabdiel Boylston scratched the skin on his six-year-old son and 285 other people and rubbed pus from smallpox scabs into the wounds. All but six of his patients survived.
>
> *Item 1:* What hypothesis might Zabdiel Boylston have been testing?
>
> *Item 2:* Give two other pieces of information that you would need to decide how successful Boylston’s approach was.

The items assess the first two of five process skills that have been identified in the framework for the assessment of scientific literacy:

(1) recognising questions that could be or are being answered in a scientific investigation;
(2) identifying evidence/data needed to test an explanation or explore an issue;
(3) critically evaluating conclusions in relation to available scientific evidence/data;
(4) communicating to others valid conclusions from available evidence/data; and
(5) demonstrating understanding of science concepts.

Knowledge of human biology and health is also required, for it is the combination of the process skill and scientific understanding that is assessed in the PISA surveys, indicated explicitly in the definition of scientific literacy (see p. 131). Thus the results will be expressed in terms of a scale of scientific literacy with a progression from less developed to more developed scientific literacy, outlined as follows:

> the student with less developed scientific literacy might be able to identify some of the evidence that is relevant to evaluating a claim or supporting an argument or might be able to give a more complete evaluation in relation

> to simple and familiar situations. A more developed scientific literacy will show in giving more complete answers and being able to use knowledge and to evaluate claims in relation to evidence in less familiar situations. (OECD, 1999, in press)

Conclusions

It has been argued that assessing science process skills is important for formative, summative and monitoring purposes. The case for this is essentially that the mental and physical skills that are described as science process skills have a central part in learning with understanding. Thus they are important in the development of 'big ideas' that are needed to make sense of the scientific aspects of the world and so must be actively developed as part of formal education. They must be included, therefore, in formative assessment.

Further, it is widely acknowledged that learning does not end with formal education but has to be continued throughout life, requiring, *inter alia*, skills of finding, evaluating and interpreting evidence. Thus the level of these skills that students have achieved as a result of their formal education is an important measure of their preparation for future life and so must be a part of summative assessment. It follows that it is important for nations and states to monitor the extent to which education systems and curricula support process skills development. In this context, a major limitation of the IEA studies has been the lack of attention in its written tests to anything other than the application of knowledge. The OECD PISA surveys, planned to begin in 2000, aim to shift the balance to processes, by assessing the ability to draw evidence-based conclusions using scientific knowledge.

The fact that process skills have to be used, and therefore assessed, in relation to some specific content has been noted as an obstacle to arriving at a reliable assessment of skills development. However, the extent to which this is a problem depends on the purpose of the assessment. For formative assessment, where the whole purpose is to use the information gained to help learning, the variation in performance due to the content can be a guide to taking action. The fact that a child can do something in one context but apparently not in another is a positive advantage, since it provides clues to the conditions which seem to favour better performance and those which seem to inhibit it (Harlen, 1996).

For formative purposes the reliability of the assessment is less important than its validity; that is, that it really does reflect the use of skills. For summative purposes reliability assumes a greater importance because the results may be used for reporting progress or comparing students and so must mean the same for all students, whoever assesses them. Unless appropriate steps are taken, the push to increase reliability can infringe validity by a preference for easily marked questions, such as those in multiple-choice formats. Skills concerned with planning investigations, criticising given procedures or evaluating and interpreting evidence cannot be assessed in this way (cf. Pravalpruk in this issue). Assessment of these skills requires more extended tasks, such as students encounter, or should encounter, in their regular work and which can be observed and marked by the teacher. The provision of such tasks may be a spin-off from the national and international monitoring programmes in science, which attract resources for developing innovative assessment tasks. What is needed is a similar investment in the development of such tasks for use by teachers, accompanied by professional development to increase the reliability of the results and confidence in teachers' judgements. Without this there will continue to be a mismatch between what our students need from their science education and what is assessed and so taught.

References

ADEY, P. (1997) It all depends on the context, doesn't it? Searching for general, educable dragons, *Studies in Science Education*, 29, pp. 45–92.

ADEY, P. & SHAYER, M. (1990) Accelerating the development of formal thinking in high school students, *Journal of Research in Science Teaching*, pp. 267–285.

ADEY, P. & SHAYER, M. (1993) An exploration of long-term far-transfer effects following an extended intervention programme in the high-school science curriculum, *Cognition and Instruction*, 11, pp. 1–29.

ADEY, P. & SHAYER, M. (1994) *Really Raising Standards: cognitive intervention and academic achievement* (London, Routledge).

ALTON-LEE, A., NUTHALL, G. & PATRICK, J. (1993) Reframing classroom research: a lesson from the private world of children, *Harvard Educational Review*, 63, pp. 50–84.

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE (1993) *Benchmarks for Scientific Literacy* (New York, Oxford University Press).

BAIRD, J. R. (1998) A view of quality in teaching, in: B. J. FRASER & K. G. TOBIN (Eds) *International Handbook of Science Education*, pp. 153–168 (Dordrecht, Kluwer).

BLACK, P. J. (1993) Formative and summative assessment by teachers, *Studies in Science Education*, 21, pp. 49–97.

BLACK, P. J. & WILIAM, D. (1998a) *Inside the Black Box: raising standards through classroom assessment* (London, School of Education, King's College).

BLACK, P. J. & WILIAM, D. (1998b) Assessment and classroom learning, *Assessment in Education: principles, policy and practice*, 5, pp. 7–74.

BRIDGEMAN, B. (1992) A comparison of quantitative questions in open-ended and multiple choice format, *Journal of Educational Measurement*, 29, pp. 253–271.

BRYCE, T. G. K., MCCALL, J., MACGREGOR, J., ROBERTSON, I. J. & WESTON, R. A. J. (1983) *Techniques for Assessing Practical Skills in Foundation Science (TAPS 2)* (London, Heinemann).

BRYCE, T. G. K., MCCALL, J., MACGREGOR, J., ROBERTSON, I. J. & WESTON, R. A. J. (1988) *Techniques for Assessing Process Skills in Practical Science (TAPS 2)* (London, Heinemann).

BRYCE, T. G. K., MCCALL, J., MACGREGOR, J., ROBERTSON, I. J. & WESTON, R. A. J. (1991) *How to Assess Open-ended Practical Investigations in Biology, Chemistry and Physics (TAPS 3)* (London, Heinemann).

CROOKS, T. J. (1988) The impact of classroom evaluation practices on students, *Review of Educational Research*, 58, pp. 438–481.

CROOKS, T. J. & FLOCKTON, L. (1996) *Science Assessment Results 1995* (Dunedin, Educational Assessment Research Unit, University of Otago).

DAUGHERTY, R. D. (1997) National Curriculum assessment: the experience of England and Wales, *Educational Administration Quarterly*, 33, pp. 198–218.

DES (1982) *Science in Schools. Age 13: report no. 1* (London, HMSO).

DES (1984a) *Science in Schools. Age 11: report no. 3* (London, DES).

DES (1984b) *Science in Schools. Age 13: report no. 2* (London, DES).

DES (1988) *Science at Age 11. A Review of APU Survey Findings from 1980–84* (London, HMSO).

DEPARTMENT FOR EDUCATION (1995) *Science in the National Curriculum 1995* (London, HMSO).

DONALDSON, M. (1978) *Children's Minds* (Glasgow, Fontana).

FENSHAM, P. (1998) Student responses to the TIMSS test, *Research in Science Education* (in press).

FENSHAM, P. J. & HASLAM, F. (1998) Conceptions of context in extended response items in TIMSS, paper presented at the Australian Science Education Research Association conference, Darwin, July.

GOTT, R. & DUGGAN, S. (1994) *Investigative Work in the Science Curriculum* (Buckingham, Open University Press).

GOTT, R., WELFORD, G. & FOULDS, K. (1988) *The Assessment of Practical Work in Science* (Oxford, Blackwell).

HARLEN, W. (1994) Towards quality in assessment, in: W. HARLEN (Ed.) *Enhancing Quality in Assessment*, pp. 139–145 (London, Paul Chapman).

HARLEN, W. (1996) *The Teaching of Science in Primary Schools*, second edition (London, David Fulton).

HARLEN, W. & JAMES, M. J. (1997) Assessment and learning: differences and relationships between formative and summative assessment, *Assessment in Education: principles, policy and practice*, 4, pp. 365–380.

HARLEN, W. & JELLY, S. J. (1997) *Developing Science in the Primary Classroom* (London, Longman).

HARLEN, W., HOLROYD, C. & BYRNE, M. (1995) *Confidence and Understanding in Teaching Science and Technology in Primary Schools* (Edinburgh, Scottish Council for Research in Education).

HEIN, G. F. (1991) Active assessment for active science, in: V. PERRONE (Ed.) *Expanding Student Assessment*, pp. 106–131 (Alexandria, VA, Association for Supervision and Curriculum Development).

HEIN, G. F. & PRICE, S. (1994) *Active Assessment for Active Science* (Portsmouth, NH, Heinemann).

HODSON, D. (1992) Assessment of practical work: some considerations in philosophy of science, *Science and Education*, 1, pp. 115–144.

HOFSTEIN, A. & LUNETTA, V. N. (1982) The role of the laboratory in science teaching: neglected aspects of research, *Review of Educational Research*, 52, pp. 201–217.

JOHNSON, S. (1989) *National Assessment: the APU science approach* (London, HMSO).

JONES, A., SIMON, S. A., BOLACK, P. J., FAIRBROTHER, R. W. & WATSON, J. R. (1992) *Open Work in School Science: development of investigations in schools* (Hatfield, Association for Science Education).

LOCK, R. (1990) Assessment of practical skills, part 2. Context dependency and construct validity, *Research in Scientific and Technological Education*, 8, pp. 35–52.

MARIK, E. A., EUBANKS, C. & GALLAGHER, T. H. (1990) Teachers' understanding and use of the learning cycle, *Journal of Research in Science Teaching*, 27, pp. 821–834.

NATIONAL ACADEMY OF SCIENCES (1994) *National Science Education Standards* (Washington, DC, National Academy of Sciences).

NUNES, T., SHLIEMANN, A. & CARRAHER, D. (1993) *Street Mathematics and School Mathematics* (Cambridge, Cambridge University Press).

OECD (1999, in press) Performance Indicators for Student Achievement (PISA), *Science Framework* (Paris, OECD).

POLLARD, A., BROADFOOT, P., CROLL, P., OSBORN, M. & ABOTT, D. (1994) *Changing English Primary Schools? Impact of the Education Reform Act at Key Stage One* (London, Cassell).

RAIZEN, S. A., BARON, J. B., CHAMPAGNE, A. B., HAERTEL, E., MULLIS, I. V. S. & OAKES, J. (1989) *Assessment in Science Education: the middle years* (Washington, DC, National Center for Improving Science Education).

RAIZEN, S. A., BARON, J. B., CHAMPAGNE, A. B., HAERTEL, E., MULLIS, I. V. S. & OAKES, J. (1990) *Assessment in Elementary School Science* (Washington, DC, National Center for Improving Science Education).

ROBITAILLE, D. F., SCHMIDT, W. H., RAIZEN, S., MCKNIGHT, C., BRITTON, E. & NICOL, C. (1993)

Curriculum Frameworks for Mathematics and Science. Third International Mathematics and Science Study Monograph No 1 (Vancouver, Pacific Educational Press).

ROTH, W.-M. (1998) Teaching and learning as everyday activity, in: B. J. FRASER & K. G. TOBIN (Eds) *International Handbook of Science Education*, pp. 169–182 (Dordrecht, Kluwer).

RUSSELL, T. J. & HARLEN, W. (1990) *Assessing Science in the Primary Classroom: practical tasks* (London, Paul Chapman).

SCHILLING, M., HARGREAVES, L., HARLEN, W. WITH RUSSELL, T. J. (1990) *Assessing Science in the Primary Classroom: written tasks* (London, Paul Chapman).

SHAYER, M. & ADEY, P. (1993) Accelerating the development of formal thinking in middle and high school students IV: three years on after a two year intervention, *Journal of Research in Science Teaching*, 30, pp. 351–366.

SWAIN, J. R. L. (1991) The nature and assessment of scientific explorations in the classroom, *School Science Review*, 72(260), pp. 65–76.

TAMIR, P. (1990) Justifying the selection of answers in multiple-choice questions, *International Journal of Science Education*, 12, pp. 563–573.

TOBIN, K., KAHLE , J. B. & FRASER, B. J. (Eds) (1990) *Windows into Science Classrooms: problems associated with higher-level learning* (London, Falmer Press).

TORRANCE, H. (Ed.) (1995) *Evaluating Authentic Assessment* (Buckingham, Open University Press).

WATSON, P. & JOHNSON-LAIRD, P. (1972) *Psychology of Reasoning: structure and content* (London, Batsford).

WELLINGTON, J. (Ed.) (1989) *Skills and Processes in Science Education: a critical analysis* (London, Routledge).

WILIAM, D. (1998) Enculturating learners into communities of practice: raising achievement through classroom assessment, paper given at European Conference on Educational Research, Ljubjana, September.

WOOLNOUGH, B. E. (Ed.) (1991) *Practical Science* (Milton Keynes, Open University Press).