Assessment & Inquiry-Based Science Education:

Issues in Policy and Practice

Wynne Harlen

Editorial Committee:
Derek Bell, Jens Dolin, Pierre Léna, Shelley Peers,
Xavier Person, Patricia Rowell and Edith Saltiel

Global Network of Science Academies (IAP) Science Education Programme

Assessment & Inquiry-Based Science Education: Issues in Policy and Practice

Wynne Harlen

In recent years there has been a rapid expansion of interest in inquiry-based science education (IBSE). Classroom and laboratory practices and materials which encourage students to take an active part in making sense of events and phenomena in the world around are being promoted and developed through pilot projects in countries across the globe. Embracing inquiry-based education recognises its potential to enable students to develop the concepts, competences, attitudes and interests needed by everyone for life in societies increasingly dependent on applications of science. Inquiry-based education also engenders reflection on the thinking processes and learning strategies that are necessary for continued learning throughout life. There are, however, many challenges in implementing IBSE. Central among these is the assessment of students' learning since this has a strong influence on what is taught and how it is taught.
This book aims to ensure that assessment practices reflect the principles and goals of IBSE and serve to help as well as to report learning.

ISBN 978-1-291-33214-8
90000
9 781291 332148

Assessment & Inquiry-Based Science Education:

Issues in Policy and Practice

Wynne Harlen

Editorial Committee:

Derek Bell, Jens Dolin, Pierre Léna, Shelley Peers, Xavier Person, Patricia Rowell and Edith Saltiel

Published by the Global Network of Science Academies (IAP) Science Education Programme (SEP)

www.interacademies.net/activities/projects/12250.aspx

TWAS-Strada Costiera, 11-34151, Trieste, Italy

Page layout by Juliet Miller

ISBN: 978-1-291-33214-8

Assessment & Inquiry-Based Science Education:

Issues in Policy and Practice

Contents

Tables

Boxes

Introduction

In recent years there has been a rapid expansion of interest in inquiry-based science education (IBSE). Classroom and laboratory practices and materials which encourage students to take an active part in making sense of events and phenomena in the world around are being promoted and developed through pilot projects in countries across the globe. Embracing inquiry-based education recognises its potential to enable students to develop the understandings, competences, attitudes and interests needed by everyone for life in societies increasingly dependent on applications of science. Inquiry leads to knowledge of the particular objects or phenomena investigated, but more importantly, it helps to build broad concepts that have wide explanatory power, enabling new objects or events to be understood. It also engenders reflection on the thinking processes and learning strategies that are necessary for continued learning throughout life. There are, however, many challenges in implementing IBSE. Central among these is the assessment of students' learning since this has a strong influence on what is taught and how it is taught.

Background to the book

The decision to write this book was stimulated by the international conference held in Helsinki, 30 May to 1 June 2012. The conference – jointly planned by the Global Network of Science Academies (IAP), ALLEA (All European Academies), the Finnish Academy of Science and Letters and Finland's Science Education Centre (LUMA) – was entitled Developing Inquiry-Based Science Education: New Issues. The issues concerned the roles of assessment in IBSE and the relationship of IBSE with industry. The greater part of the conference was concerned with issues relating to assessment, which is the focus of this book.

The Science Education Programme (SEP) of the IAP, under the leadership of its founder, Professor Jorge Allende, held major international conferences on various aspects of IBSE in 2005, 2008 and 2010, interspersed with meetings of working groups to plan and report on activities. The project has also sponsored activities through four Regional Networks of academies across the world. Following the retirement of Professor Allende, the SEP has been led by Professor Pierre Lena since 2011.

Throughout the work of the IAP SEP it has been recognised that the role of student assessment has become increasingly important in the understanding and implementation of IBSE. Assessment is a key aspect of strategic planning for education change. Accompanying the spread of pilot projects there has been a demand for information about the effectiveness of IBSE in order to justify the resources needed for its implementation. Good measures of the outcomes of IBSE are needed for this. But providing information about students' achievement is only one role of assessment; its role in helping learning and developing deeper understanding relating to the goals of science education has gained considerable support in recent years throughout the world. However, the acknowledged influence of what is assessed on curriculum content and pedagogy means that assessment can also restrain change in science education when established assessment methods and content do not reflect the goals of IBSE.

Addressing the issues that inhibit the implementation of IBSE becomes more important in view of the value of inquiry-based learning in developing those skills which are required by the current and future work force. For these reasons the Global Council of the IAP SEP at its meeting in Paris in April 2011 decided that student assessment would be a major focus of its 2012 conference and led to the decision to produce this book.

Box 1: Countries represented at the Helsinki conference

Argentina
Australia
Austria
Brazil
Cameroon
Canada
China
Colombia
Costa Rica
Denmark
Estonia
Finland
France
Germany
Ghana
Haiti
Hungary
India
Iran
Israel
Italy
Mauritius
Mexico
Kenya
Kosova
Malaysia
Montenegro
Mozambique
Netherlands
New Zealand
Nigeria
Norway
Pakistan
Senegal
Serbia
Slovenia
South Africa
Sudan
Sweden
Switzerland
Tanzania
Tunisia
Uganda
Ukraine
United Kingdom
United States of America
Venezuela
Vietnam
Zambia
Zimbabwe

Major points emerging from the conference

The presentations at the Helsinki conference, and particularly the discussions among the 93 participants from 50 countries (see Box 1), indicated major concerns relating to student assessment, briefly outlined here under six headings:

- The need for clarification of terms
- Understanding the interactions between assessment, pedagogy, curriculum content and education policy frameworks
- The role of tests, tasks and teachers in assessing IBSE
- Fitness of assessment for different purposes
- Challenges of making changes in assessment
- The need for more research relating to assessment.

The need for clarification of terms

If we are to make progress in developing policy and practice in relation to the assessment of IBSE it is clearly important for there to be a shared understanding of the words used. Discussion at the conference uncovered some confusion about the meaning of key terms such as assessment, testing, and evaluation, and about the difference between formative and summative assessment. Language differences may be a problem here, particularly where a language has a single word to mean both assessment and evaluation. It also appeared that the meaning of IBSE continued to be rather narrowly interpreted as concerned only with the development of skills. This is understandable given that IBSE has been introduced in some countries as an antidote to textbook-based teaching. However, having a common view of what IBSE means in practice and how it differs from other approaches to teaching and learning is not only necessary for implementation but essential as a platform for developing assessment. Conference participants pointed out that, apart from the meanings of words, there are cultural differences underlying the discourse about IBSE and its assessment, leading to different systemic visions which need to be recognised and debated.

Understanding the interactions between assessment, pedagogy, the curriculum content and education policy frameworks

There was a general consensus that what is assessed influences the priority given by teachers to various goals of learning and therefore it is essential that all important goals are included in what is assessed. It was suggested that this would be facilitated if the curriculum and assessment were constructed together by the same agencies, to avoid inspirational goals being effectively converted into a series of disconnected tasks. A related point, made by several participants, was that assessment should be consistent with the theory of learning that underpins IBSE. There was a strong opinion that the connection between good assessment, the implementation of IBSE and the development of key scientific ideas needs to be spelled out. A description is needed of what is meant by 'quality' in science learning and how this is different at different ages as students progress in their learning 'of science' and 'about science'. Assessment should support, and be seen to support, the development of good citizenship and of the knowledge and skills needed to tackle major global problems. Educational policies based on ambitions for high levels of performance in tests and international surveys need to be reconciled with concerns for high quality education for all.

The role of tests, tasks and teachers in assessing IBSE

The conference participants recognised that the nature of IBSE poses many conceptual, logistic and technical challenges for student assessment. Conceptually, assessment procedures should enable an understanding of what it means to say someone has learned science well. This implies clarity about the goals of science education and how IBSE contributes to them. Technically, the challenges are to ensure reliability without compromising validity. The agreed aims of developing confident, autonomous and collaborative learners are hard to assess directly and surrogate measures reflecting these qualities have to be sought. It was recognised that it is almost impossible for tests of a reasonable length that can be taken by all students to provide the rich information needed to assess IBSE goals. Furthermore, producing and administering tests incurs a high cost in time as well as other resources. Participants affirmed that teachers should be involved in conducting assessment at some or all stages of schooling. It was suggested that feedback on the outcomes of assessment and teacher appraisal could have a role in strengthening teachers' assessment skills. The impact of school culture has also to be recognised. Teachers may be constrained in adopting new approaches to teaching and assessment by accountability measures which take a narrow view of teaching and learning science.

Fitness of assessment for different purposes

For assessment to be used to help learning means that teachers incorporate formative assessment strategies as part of their pedagogy rather than adding a series of mini-summative assessment events. For summative assessment, tests are commonly used for checking performance at the end of topics or courses and for producing reports on progress at regular intervals. It was widely agreed that most science tests used by teachers do not reflect the goals of IBSE. Since tests have a strong influence on what is taught, it is important to consider alternatives to most current forms of test in order to obtain more dependable information about the learning that results from IBSE. Some conference participants expressed the view that using test results for the purposes of accountability inhibits learning. It was noted that large scale test programmes such as TIMSS and PISA use methods that do not support valid inferences about performance in IBSE at the system level. However, it was recognised that such survey results do have a role but should be regarded as just one indicator of system performance, to be considered within a framework of information about other aspects.

Challenges of making changes in assessment policy and practice

Some of the greatest changes and the greatest challenges to change were perceived by conference participants as being in relation to involving teachers in assessment. They need help to develop 'assessment literacy'. Several goals of IBSE are better assessed by teachers. The classroom opportunities that have to be provided for students to learn through IBSE are equally opportunities for teachers to assess and record their students' developing understanding and skills. Teachers also need tools to support their formative and summative assessment. In many countries there are challenges in relation to participation of girls in science, but there is no evidence of gender differences in competence in IBSE. It is important to influence pre-service course providers so that teachers enter the profession aware of IBSE and how it should be included in their assessment practice. Conference delegates also advised that politicians, administrators, parents and the public in general need to be educated about the meanings, purposes, strengths and limitations of assessment.

The need for more research relating to assessment

Research reported in conference presentations provided striking evidence of the need to rethink certain aspects of how students are assessed if this is to give them opportunities to show what they know and can do. Confidence in the accuracy of some assessment results is often shaken by research showing that how students approach a question or task and what processes of thinking they use in tackling it are not what was intended and assumed to be the case when results are interpreted. More such studies are needed in order to improve the validity of assessment tools and procedures and to deter unrealistic assumptions of accuracy. It is known that even minor changes in format and working of written questions can influence students' performance but less is known about the differential impact on students of different gender, experience, and background. Several other areas where research is needed were mentioned. These include: the impact of alternatives to tests; the combination of different types of evidence that provides the best picture of certain kinds of student performance; how trust in teacher-based assessment can be increased; how to train teachers in both the formative and summative uses of assessment; how to combine formative and summative assessment in national and district assessment systems.

The focus of this book

In this book, as in the conference, we have distinguished between two sets of issues, concerning:

- the processes and uses of *assessment of students' learning*, and
- the *value* of inquiry-based science education and the *evaluation of its effectiveness*.

The first of these relates to the various issues, summarised above, that arise from recognising the role student assessment plays as an integral part of students' learning experiences. Assessment was once regarded as something that takes place after learning and as being quite separate from the process of learning. This view is no longer tenable; assessment is now acknowledged as a central part of education, with a proven role in helping learning as well as in reporting it. How the results of student assessment are used is recognised as having an important influence, which can be positive or negative, on the content and methods of teaching. Thus the nature of the assessment, and particularly the extent to which it allows students to show what they know and can do in relation to intended learning goals, are key factors in students' education. In the context of IBSE it is a matter of concern that most current assessment tools and procedures fall short of what is needed to provide a good account of students' achievement of the goals of IBSE.

In relation to the second of these sets of issues, the value of IBSE is not a matter that can be decided by empirical evidence, but is a value judgement that the competences, understanding, interests and attitudes that are its aims are worthwhile and indeed are necessary in a modern education. What programme evaluation *can* show is the extent to which students achieve these aims through experiences designed to implement IBSE. Assessment of students has a role in programme evaluations, but there are many other factors involved. In particular what students achieve is only informative in relation to IBSE outcomes if there is evidence that students are truly learning through inquiry and that the data provided by the assessment enable inferences to be drawn about the scientific understanding and science inquiry skills that are the aims of IBSE. The nature of the assessment of students' learning in science is one of the main factors holding back the implementation of IBSE programmes.

There is undoubtedly a need to answer questions about the impact of inquiry-based education on students' achievement, but this cannot be done without valid tools. Thus by focusing here on the first set of issues above – concerning the assessment of students' learning – we not only offer the promise of providing some of the necessary tools for programme evaluation, but, more importantly, consider assessment as an integral part of students' learning through inquiry.

Chapter 1
Clarification of terms

Many of the terms used in discussing assessment have both technical and commonplace usage, as is the case of many terms used in science. Lack of clarity in meaning and consistency in usage impedes good communication and the understanding of distinctions and connections between concepts. While not everyone will agree on the definition of the terms used in assessment – and particular problems are associated with translation into other languages – it is important at least to make explicit the meanings being given to words used in this book, in particular 'assessment', 'testing', 'evaluation' and associated words such as standards, criteria, validity and reliability. The meanings of inquiry are discussed in Chapter 2 and in Chapter 3 we look at other terms relating to the purposes, functions and uses of assessment.

Assessment, evaluation and appraisal

The OECD, in its reviews of evaluation and assessment in its member countries, makes a clear and useful distinction between student assessment, teacher appraisal, school evaluation and system evaluation, at the same time as recognising that these system components need to work together in policies that aim to improve learning outcomes:

> The term "assessment" is used to refer to judgements on individual student performance and achievement of learning goals. It covers classroom-based assessment as well as large-scale, external tests and examinations. The term "appraisal" is used to refer to judgements on the performance of school-level professionals, e.g. teachers and principals. Finally, the term "evaluation" is used to refer to judgements on the effectiveness of schools, school systems and policies.[1]

Assessment and evaluation both describe a process of generating and interpreting evidence for some purpose. They both involve decisions about what evidence to use, the generation and collection of that evidence in a systematic and planned way, the interpretation of the evidence to produce a judgement, and the communication and use of the judgement. It is worth noting that the evidence, of whatever kind, is only ever an indication or sample of a wider range that could be used.

In this book, following the OECD convention, the word 'assessment' is used to refer to the process of collecting and using evidence about the outcomes of learning, usually students' learning, but it can also refer to the learning of others such as teachers.

'Evaluation' is used in relation to generating and using evidence about systems, materials, procedures and processes. Evaluation of schools, systems and teaching approaches may make use of evidence of students' learning, but the judgement is about the value or success of other things such as school policies and programmes rather than the learning of students, although this may be part of the evidence used in the evaluation.

[1] Nusche, D. et al (2012) *OECD Reviews of Evaluation and Assessment in Education: New Zealand 201*1. OECD Publishing, p24

Testing and other methods of assessment

Although the terms assessment and testing are sometimes used interchangeably there is an important distinction between them. Testing may be regarded as a method of collecting data for assessment, thus assessment is a broader term, covering other methods of gathering and interpreting data as well as testing.

A closer look at what assessment involves helps to clarify this relationship and to identify other aspects of assessment involving words such as 'standards' and 'criteria'.

All assessment of students' achievements involves the generation, interpretation, communication and use of data for some purpose. In just this simple statement there is room for an enormous range of different kinds of activity, but each will involve a) students being engaged in some activity, b) the collection of data from that activity by some agent, c) the judgement of the data by comparing them with some standard and d) some means of describing and communicating the judgement. There are several forms that each of the components of assessment can take.

a) Activities in which students are engaged can be, for example:

- their regular work
- some written or practical tasks created by the teacher for the purpose of assessment
- some written or practical tasks created externally.

b) The data can be collected by:

- the teacher
- the students
- the teacher and students together
- an external agent (examination board, qualifications authority, test developer).

c) The data can be judged in relation to:

- norms, in which the standard of comparison is the performance of other students (norm-referenced)
- criteria, in which the standard of comparison is a description of aspects of performance (criterion-referenced)
- students' previous performance, in which an individual's performance is judged in relation to the student's other or earlier performance (student-referenced or ipsative).

d) The judgements can be communicated as:

- a written or oral comment by the teacher
- a mark or score or percentage
- a profile of achievement
- a level or grade
- a ranking or percentile.

Different assessment tools and procedures are created by different combinations of these various ways of collecting, judging and communicating data. For example, a standardised test comprises tasks created by an external agency which will have trialled the test during development with a large sample of the appropriate population, so that an individual's score can be expressed in terms of comparison with the 'norm' for that population. The result will indicate whether a student's performance is above or below average but not what he or she can do.

A criterion-referenced test differs from a norm-referenced test by being designed to give information about what a student can do in relation to specified outcomes. The items will be chosen for their relevance to the curriculum so that the results can be used in establishing, not how a student compares with others, but how his or her performance compares with the intended performance. At the same time, the target level of performance is set by reference to what can be expected of the population for whom the test is intended. Thus there is a normative element in deciding the criteria against which performance is judged.[2] When tests are used the data are restricted to the items in the test, whereas if the assessment is carried out by teachers there is the potential to use the full range of learning activities in making judgements using criteria of expected performance in relation to lesson goals.

Validity, reliability, resources and manageability

To decide the best way of conducting assessment in a particular case it is necessary to consider the properties of possible tools in relation to the purposes and uses intended for the assessment results. One obvious desirable property is that any assessment should be valid for its purpose; that it assesses what it is intended to assess. Another is that it should provide reliable, or trustworthy, data. But there are also other matters to be taken into account; in particular, in view of the interdependence of the various system components, the impact on other assessment practices, on the curriculum and on pedagogy. Further, the use of resources – assessment can be costly, both in terms of monetary resources and the time of students and teachers – and manageability need to be considered.

Validity

It is usual to define validity of an assessment in terms of how well what is assessed corresponds with the behaviour or learning outcomes that it is intended should be assessed. Various types of validity have been proposed depending on the kind of information used in judging the validity. For instance, *content validity* refers to how adequately the assessment covers the subject domain being taught and is usually based on the judgement of experts in the subject. However content coverage is not enough to distinguish a test or other assessment of science learned through inquiry from science learned in other ways. *Construct validity* is a broader concept, reflecting the full range of outcomes of learning in a particular subject domain. The important requirement is that the assessment samples all aspects – but only those aspects – of students' achievement relevant to the particular purpose of the assessment. Including irrelevant aspects is as much a threat to validity as omitting relevant aspects.

However, this view of validity as a property of an assessment method or instrument, regardless of the circumstances in which is it used and the uses that are made of the results, has been widely challenged. Newton[3] has pointed out the error in both these assumptions. In the case of the conditions of use of the assessment, the accuracy of the results as a measure of the construct will depend on how the assessment is administered as well as what it contains. In the case of the use of the results, it may be that claims are made about what was intended to be assessed when in fact other factors were more influential in the results (as when a test of mathematics has a high reading demand which makes it uncertain whether it is reading or mathematical ability that is most influential in the results.)

The notion of validity that takes into account not just how well the assessment samples the construct it is intended to assess but what is claimed on the basis of the results, is one that relates to the inferences drawn from the results. It was formally expressed in a widely quoted definition of validity by Messick:

> Validity is an integrative evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.[4]

[2] Black, P (1998) *Testing: Friend or Foe?* London: Falmer Press.

[3] Newton, P. Validity, purpose and the recycling of results. In (ed) J. Gardner *Assessment and Learning*. 2nd edn. London: Sage

[4] Messick, S.(1989) Validity, in (ed) R. Linn *Educational Measurement* (3rd edn)American Council on Education , Washington: Macmillan, pp 13-103, p13.

A consequence of adopting this view of validity is to recognise that validity can vary due to factors affecting performance such as conditions of testing, which are usually described in terms of the reliability of the assessment. We return to this point after considering the meaning of reliability.

Reliability

The reliability of an assessment refers to the extent to which the results can be said to be of acceptable consistency or accuracy for a particular use. This may not be the case if, for instance, the results are influenced by who conducts the assessment or they depend on the particular occasion or circumstances at a certain time. Thus reliability is often defined as, and measured by, the extent to which the assessment, if repeated, would give the same result.

Reliability has meaning mostly in the case of summative assessment and particularly for tests. When assessment is used formatively (see Chapter 3), it involves only the students and the teachers and the notion of making a repeatable judgement and treating all students in the same way is not relevant. No judgement of grade or level is involved; only the judgement of how to help a student take the next steps in learning, so reliability in this formal sense is not an issue. For formative assessment what is important is 'the quality of information that is gathered and provided in feedback'.[5]

Box 2: The reliability/validity trade-off

In any assessment there is a limit to the extent that both reliability and validity can be optimised. This applies to whatever form the assessment takes, but it is most readily identified in relation to using tests. In selecting items for a test, in order to increase reliability, it is inevitable that the selection gives preference to those items that can be consistently marked or marked by machine. This favours items assessing factual knowledge and the use of a closed item format, as opposed to items requiring application of knowledge and the use of more open-ended tasks. The consequent limitation on what is covered in a test affects its validity. Attempts to increase validity by widening the range of items, say by including more open-response items where more judgement is needed in marking, will mean that the reliability is reduced. Thus there is a trade-off between reliability and validity; increasing reliability decreases validity and vice versa.

However, high reliability *is* necessary when the results are used by others and when students are being compared or selected. Thus the discussion here relates to tests, where reliability is important and it is assumed that all students will experience the same conditions and have the same opportunities. Of course it is never the case that the same conditions mean the same opportunities for different students, or that an individual student reacts in the same way to a test on different occasions. Individual students' responses to the same conditions will vary from day to day and, most importantly, the particular selection of items in a test will be found more difficult by some students than others of equal ability. All test items present questions in some context and there is research evidence that students who perform well in one item will not necessarily do so in another testing the same concepts or skills but set in a different context. For any test there is a large number of possible items and only a small sample of them can be included in a test of a reasonable length. A different selection would produce a different result, giving rise to what is described as the 'sampling error'.

[5] Stobart, G. (2012) Validity in formative assessment In (ed) J. Gardner *Assessment and Learning.* 2nd edn. London: Sage p234.

The sampling error can be much larger than is generally realised. For example, Wiliam[6] has estimated that for national tests in England about 40% of students will be assigned to the 'wrong' grade level, even though these levels each span roughly two years. A way of reducing this source of error would be to increase the number of contexts included for each competence assessed and thus the number of items used. But the length of a test cannot be greatly increased without incurring other forms of error (student fatigue, for instance) so more items per skills or concept would mean fewer skills and concepts included, thus reducing the range of what is assessed and so reducing the validity of the test. This is a further example of the interaction between reliability and validity identified in Box 2. The consequence is greatest in relation to individual student testing where all have to be given the same items. The effect is much less in relation to population surveys where the sampling error can be reduced by using a large number of items spread across equivalent samples of students who take different groups of items (see Chapter 6).

Resources and manageability

The resources required to provide an assessment ought to be commensurate with the value of the information for users of the data. The resources may be teachers' time, expertise and the cost both to the school and to external bodies involved in the assessment. In general there has to be a compromise, particularly where a high degree of accuracy is required. There is a limit to the time and expertise that can be used in developing and operating, for example, a highly reliable external test or examination. Triple marking of all test papers would clearly bring greater confidence in the results; observers visiting all candidates would increase the range of outcomes that can be assessed externally; training all teachers to be expert assessors would have great advantages – but all of these are unrealistic in practice. Balancing costs and benefits raises issues of values as well as of technical possibilities.

The cost of formative assessment is negligible once it is incorporated into practice. The process of introducing it may well be considerable in terms of teachers' time for professional development. Good formative assessment, as discussed in Chapter 5, requires not only mastery of certain classroom strategies but knowledge of routes of progression in aspects of learning and examples of teachers and students using evidence to identify next steps in learning. These costs, however, are integral to efforts to improve learning.

Summative assessment requires resources in terms both of teachers' and students' time. When tests developed by agencies outside the school or by commercial publishers are used, there is considerable cost. Even if national tests and examinations are provided free to schools, the cost has to be borne by the system and can be surprisingly large. If the direct costs of producing, distributing, scoring, tests and so on, are added to the time taken up by preparing for and taking external tests and examinations, the total can amount to a significant proportion of the education budget.[7] It certainly constitutes a case for considering the balance between costs and benefits in deciding the methods to be used for summative assessment.

[6] Wiliam, D. (2001) Reliability, validity and all that jazz, *Education* 3-13, 29 (3): 17-21.

[7] Harlen, W. (2007) *Assessment of Learning*. London: Sage p 61/2

Chapter 2
Inquiry-based science education: rationale and goals

Decisions about the most effective ways of ensuring that assessment is used to support and report science learning through inquiry require clear identification of the intended learning outcomes from IBSE. The assessment approaches that provide the most dependable data about these outcomes can then be selected or devised. In this chapter we take a close look at the goals of IBSE and the reason for the importance of these goals. Before embarking on this, however, it is important to note that inquiry is not the only approach used in science education. There are aspects of learning science, such as knowledge of scientific vocabulary, conventions and use of equipment, that are best learned through direct instruction. Thus not all science teaching and not all assessment will be concerned with the specific outcomes of learning through inquiry. However, knowledge of facts and procedures are means to the end of developing understanding through inquiry, thus the major element in assessment should reflect the understanding, skills and competences that are the goals of IBSE.

Inquiry in science education

Inquiry is a term used both within education and in daily life to refer to seeking explanations or information by asking questions. It is sometimes equated with research, investigation, or 'search for truth'. Within education, inquiry can be applied in several subject domains, such as history, geography, the arts, as well as science, mathematics, technology and engineering, when questions are raised, evidence is gathered and possible explanations are considered. In each area different kinds of knowledge and understanding emerge. What distinguishes *scientific inquiry* is that it leads to knowledge and understanding of the natural and made world through direct interaction with the world and through the generation and collection of data for use as evidence in supporting explanations of phenomena and events.

Inquiry is by no means a new concept in education, being based on recognition of children's active roles in developing their ideas and understanding. The studies of Piaget[8] and the arguments of Dewey[9] among others in the first half of the 20th century drew attention to the important role in their learning of children's curiosity, imagination and urge to interact and inquire. More recently the US National Research Council spelled out the value of students engaging in making observations, posing questions, using tools to gather, analyse and interpret data and communicating the results.[10] Similarly the US National Science Foundation defined inquiry teaching as leading 'students to build their understanding of fundamental scientific ideas through direct experience with materials, by consulting books, other resources, and experts, and through argument and debate among themselves.'[11]

In the course of various pilot projects the past decade, the IAP Science Education Programme has formulated this definition of inquiry-based science education:

[8] Piaget, J (1929) *The Child's Conception of the World*. New York: Harcourt Brace.

[9] Dewey, J. (1933) *How we think: A restatement of the relation of reflective thinking to the educative process*. Boston, MA: D.C. Heath

[10] National Research Council (NRC) (1996*) National Science Education Standards*. Washington DC: National Academy Press

[11] National Science Foundation (NSF) (1997) *The Challenge and Promise of K-8 Science Education Reform*. Foundations, 1. Arlington, VA: NSF p7

> IBSE means students progressively developing key scientific ideas through learning how to investigate and build their knowledge and understanding of the world around. They use skills employed by scientists such as raising questions, collecting data, reasoning and reviewing evidence in the light of what is already known, drawing conclusions and discussing results. This learning process is all supported by an inquiry-based pedagogy, where pedagogy is taken to mean not only the act of teaching but also its underpinning justifications.[12]

Some words in this definition deserve emphasis and comment.

Progressively developing key ideas underlines the importance of identifying a few overarching ideas that help us to make sense of the phenomena in the world around, and then ensuring that through their science learning activities students make progress towards developing these ideas.

Learning how to... build their knowledge and understanding implies the active role of the students in their learning, which is part of formative assessment, discussed in Chapter 3, and implies a view of learning as being constructed by learners, described in Chapter 4.

Using skills employed by scientists means in addition to those skills listed, being rigorous and honest in collecting and using sufficient and relevant data to test hypotheses or answer the questions raised. Scientists check and repeat data collection, where possible, they interpret and attempt to explain their findings. Throughout their investigations they keep careful records, and in drawing conclusions they consult related existing work and present their work to others, in writing or at conferences, and share their ideas. It is obvious in the case of scientists, but worth noting for application of inquiry in school science, that those engaged in inquiry do not know the answer to the question or problem being studied, find it important to investigate and are excited about trying to find an answer or solution.

Raising questions underlines the point that students are engaged in answering questions of real interest to them that have stimulated their curiosity. Often these questions will be raised by the teacher, other students or emerge from reading but, whatever the origin of the question, in inquiry students take them as their own, engaging their curiosity and desire to understand. Raising and answering questions is sometimes equated with problem-solving, where the focus is on finding a solution that 'works'. However, in science the single solution is not enough. Developing theories and models in order to explain phenomena requires that ideas are 'evaluated against alternative explanations and compared with evidence.... Thus knowing why the wrong answer is wrong can help secure a deeper and stronger understanding of why the right answer is right.'[13]

Discussion of definitions makes clear that learning science through inquiry is a complex process in which knowledge and understanding and skills of collecting and using evidence are linked together interactively. The skills that are essential to the building of understanding are both physical and mental skills, concerned with generating evidence and using evidence to test ideas that may help to explain an event or phenomenon being studied.

At the same time, the use of skills involves knowledge and understanding, not only knowing how to generate, collect and interpret data but also understanding why it is important to work scientifically. Further, there is an affective element to the process, influencing willingness to engage in the various actions involved in pursuing an inquiry and to take notice of results which may require a change in pre-existing ideas. All this, too, is embedded in a cultural context which can promote or inhibit the development of understanding through inquiry.

[12] IAP (2012) Taking Inquiry-Based Science Education into Secondary Education. Report of a global conference. http://www.sazu.si/files/file-147.pdf

[13] National Research Council (2012) *A Framework for K-12 Science Education*. Washington DC: National Academies Press. p44

Acknowledgement of this interdependence of knowledge and skills has led to the suggestion that inquiry is best specified in terms of *practices*, complex sets of actions that lead to experiencing and understanding science as 'a body of knowledge rooted in evidence'.[13] However, in this book we continue to use the word 'skill' and 'competence' interchangeably, following the convention of the OECD:

> In the context of the OECD Skills Strategy, the concepts of 'skill' and 'competence' are used interchangeably. By skill (or competence) we mean: the bundle of knowledge, attributes and capacities that enables an individual to successfully and consistently perform an activity or task, whether broadly or narrowly conceived, and can be built upon and extended through learning.[14]

Rationale

Inquiry-based learning is complex and is not an easy option. We strive to implement it because we believe that it promotes the understanding and development of skills needed by students to meet the demands of twenty-first century life. It is widely accepted[15] that science education should enable students to develop key science concepts (big ideas) which enable them to understand the events and phenomena of relevance in their current and future lives. Students should also develop understanding of how science ideas and knowledge are obtained and the skills and attitudes involved in seeking and using evidence.

Young people will have to make more choices than did those living in past decades. They will need to develop the skills, the will, the flexibility in thinking and the energy needed to make effective decisions. The ability to continue learning throughout life is acknowledged as essential for future generations and thus it has to be a feature in the education of students in all countries, as underlined by the OECD:

> Students cannot learn in school everything they will need to know in adult life. What they must acquire is the prerequisites for successful learning in future life. These prerequisites are of both a cognitive and a motivational nature. Students must become able to organise and regulate their own learning, to learn independently and in groups, and to overcome difficulties in the learning process. This requires them to be aware of their own thinking processes and learning strategies and methods.[16]

Furthermore there is widespread recognition of the importance of developing the skills, attitudes, knowledge and understanding that are regarded as more important than accumulating large amounts of factual knowledge. Content knowledge can be found readily from the information sources widely available through the use of computers and especially the internet. What learners need are the skills to access these sources and the understanding to select what is relevant and to make sense of it.

Learning is a social activity in which language has a key role. Interaction with others often means that individuals arrive at a shared understanding of ideas that they may not have reached alone. The ideas that students form from direct experience have to be communicated and this involves using words that convey meaning to others. The process of expressing ideas through talk or writing often means that ideas have to be reformulated in ways that are influenced by the meaning that others give to words. It is also necessary to learn that science uses words with precise meanings different from their common use in everyday language, and uses mathematics and other abstract symbols when quantifying observations of the world. We return to the role of language and particularly talk, in Chapter 5.

[14] OECD (2011) *Towards an OECD Skills Strategy*. Paris: OECD. P7 footnote

[15] OECD (2003), *The PISA 2003 Assessment Framework* Paris: OECD p132
Harlen (Ed) (2010) Principles and Big Ideas of Science Education.

[16] OECD (2000) *Measuring Student Knowledge and Skills: A new Framework for Assessment*. Paris: OECD. p90

Goals

Whilst it is important for students to learn how to learn and develop the skills of inquiry, there needs to be a balance between conceptual learning and learning about how to go about learning. Learning about how to answer a question is not enough on its own; the question also needs to be answered. On the other hand, finding the answer to a particular question is not enough, for only by attending to how it was answered will the activity help learning in new contexts.

In summary, through their science education students should develop:

- understanding of fundamental scientific ideas
- understanding of the nature of science, scientific inquiry, reasoning
- scientific competences of gathering and using evidence
- scientific attitudes, both attitudes within science and towards science
- skills that support learning throughout life
- ability to communicate using appropriate language and representations, including written, oral and mathematical language
- appreciation of the contribution of science to society and of how science is used in technology and engineering.

An inquiry approach, if carried out effectively, offers the promise of achieving these aims to a greater degree than traditional approaches to teaching and learning science.[17] The critical reservation here is that it is 'carried out effectively'. The complexity of IBSE, as noted earlier, makes this a considerable challenge. Implementation may require fundamental change in several aspects of pedagogy, from the arrangement of learning space (so that students can work collaboratively) to the questions teachers ask, the feedback they give to students and the nature of their interaction with students and students' interaction with the objects and phenomena they investigate. The degree of change that may be required by comparing the actions of students engaged in inquiry-based learning with those learning, in Box 3, with those of transmission science teaching, in Box 4.

Teaching as 'transmission of facts' was the predominant mode at the time when the main aim of science education was to provide future scientists with essential knowledge, rather than also being to provide everyone with the opportunities to achieve the goals listed above. Science was passed on to students 'ready-made', as opposed to being experienced, with knowledge being created through action. But this is not to claim that inquiry is the only form of pedagogy that students encounter in their science education. There are some things to be learned such as skills of using equipment, names, conventions and symbols which are best taught directly and there will be occasions where inquiry contributes to making sense of experience without being the sole approach used. However, when understanding is the aim, inquiry has a key role in students' science education.

[17] Minner, D.D., Levy, A. J, and Century, J. (2010) Inquiry-Based Science Instruction—What Is It and Does It Matter? Results from a Research Synthesis Years 1984 to 2002, *Journal of Research in Science Teaching*, 47 (4) 474-496

Box 3: Student activities: learning through inquiry

- Students pursue questions which they have identified as their own even if introduced by the teacher
- They do not know the answer to the questions they investigate
- They know enough about the topic to engage with the question
- They make predictions based on their emerging ideas about the topic
- They take part in planning investigations to test their predictions
- They conduct investigations themselves
- They use appropriate sources and methods of collecting data relevant to testing their predictions
- They discuss what they find in relation to their initial expectations or predictions
- They draw conclusions and try to explain what they find
- They compare their findings and conclusions with what others have found and concluded
- They keep notes and records during their work
- They engage in discussion of the methods used and the results of their investigations

Box 4: Student activities: learning through transmission methods

- Students' activities follow a sequence set out in a text-book or by the teacher with little attention to placing what they do in the context of a question they want to answer
- They may read about how to conduct investigations but have little opportunity to experience the process for themselves.
- They may observe demonstrations by the teachers but may not understand the reasons for what is being done
- When they do undertake practice activities they follow instructions taking little part in deciding what to do
- The experiments they observe or conduct are designed to confirm a conclusion already known: 'experiment to show that …'
- They do not always know why certain steps in an experiment or investigation have to be carried out
- They write accounts of investigations in a structured form, often copied from a book or dictated by the teacher
- They record the 'right answer' even if they did not observe what ought to have happened
- They work independently or in pairs and are discouraged from discussing their work

Chapter 3
Assessment purposes and uses

As stated in Chapter 1, all assessment involves the generation, collection, interpretation and communication of data. The process is similar whether the purpose is to help learning or to summarise and report it. In this chapter we explore the two main purposes of assessment – to help learning or to summarise and report it – and consider the feasibility and desirability of using data collected with one purpose in mind for another purpose. Whether a particular assessment practice is formative assessment or summative assessment is essentially determined by the *use* made of the data. So, for instance, a classroom test may be used to help students and teacher to identify what students already know about a new topic – a formative use – or to judge and report on what they have learned at the end – a summative use. Thus correctly we should refer to the 'formative use of assessment' and 'the summative use of assessment', but convention and convenience allow the shorter titles which we use here. After reviewing the nature and importance of assessment for these two purposes we consider the relationship between them. The discussion of this relationship is continued later, in Chapter 6, after reviewing some procedures for implementing formative and summative assessment.

Purposes and uses

A key question to ask of any assessment is: What is its principal purpose?[18] It is generally agreed that there are two main answers to this question:

- to help students while they are learning
- to find out what they have learned at a particular time.

These are described as formative and summative purposes of assessment.

Formative assessment has the purpose of assisting learning and for that reason is also called 'assessment *for* learning' (AfL). It involves processes of 'seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning and where they need to go and how best to get there'.[19]

Summative assessment has the purpose of summarising and reporting what has been learned at a particular time and for that reason is also called 'assessment *of* learning' (AoL). It involves processes of summing up by reviewing learning over a period of time, and/or checking-up by testing learning at a particular time.

For formative assessment there is one main use of the data – to help learning. If the information about student learning is not used in decisions intended to help that learning, then the process cannot be described as formative assessment. By contrast, the data from summative assessment can be used in several ways, some relating to individual students and some to the aggregated results of groups or populations, not all of which are appropriate or valid uses. As noted in Chapter 1, validity is not a property of a particular assessment instrument or procedure, but depends on how it is used and the inferences drawn from the results of its use. The results of a test of knowledge recall do not indicate achievement across the whole subject domain – a test of arithmetic should not be used to

[18] Stobart, G. (2008) *Testing Times. The uses and abuses of assessment*. London: Routledge

[19] Assessment Reform Group (ARG) (2002) *Assessment for Learning: 10 Principles*. www.assessment-reform-group.org

indicate achievement in mathematics, nor should it be used as a measure of quality of teaching. Obvious though these points may seem, it is a fact that such abuse of test data does occur. For example, Newton[20] identified 16 uses of the results of national tests in England. These ranged from programme evaluation, target setting for students and schools, school monitoring and student selection to school choice by parents and even the valuation of property in the areas of schools with high or low test scores. In some cases it is legitimate to use measures of student performance as part of the data used in making judgements (for instance, as an element in school evaluation), but use as a sole measure - and particularly where rewards and penalties are attached – invites inappropriate actions to inflate the measured performance. We return to this in Chapter 4 when considering the use of aggregated achievement data for evaluation of the school and system and target setting.

We now take a closer look at these two main purposes of assessment – formative and summative – in each case discussing what it is and why it is important. Chapters 5 and 6 look at methods of implementation.

Formative assessment

What it is

Some of the various definitions of formative assessment proposed over the past two decades have been reviewed by Wiliam[21] who suggests that the main features are brought together in the following definition:

> Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence what was elicited.[22]

Figure 1[23] represents the processes involved as a cycle of events. Formative assessment is not something that happens occasionally; it is integral to the process of making decisions that is happening all the time in teaching. The activities represented by A, B, and C are directed towards the goals of the lesson, or series of lessons on a topic. These goals, shared with the students by the teacher, are expressed in *specific* terms; for example in a science lesson they might be 'to plan and to carry out an investigation of the conditions preferred by woodlice'. The students' work in activity A, directed to the goals, provides opportunity for both teacher and students to obtain evidence of progress towards the goals.

In order to interpret the evidence, in this example both teacher and students need to know what 'good planning' means, so students need to have some understanding of the criteria to apply in assessing their work (Is the planned investigation taking account of all relevant variables? What and how will evidence be gathered?) The judgement leads to the decision about the relevant next steps which may be to intervene or simply to move on. As Wiliam points out, 'formative assessment need not alter instruction to be formative – it may simply confirm that the proposed course of action is indeed the most appropriate'.[24] Activity B is the result of this decision and the source of evidence in a further cycle of eliciting and interpreting evidence.

[20] Newton, P.E. op cit pp270-272

[21] Wiliam, D. (2009) An integrative summary of the research literature and implications for a new theory of formative assessment, in (eds) H. L. Andrade and G. J. Cizek, *Handbook of Formative Assessment*, New York: Taylor and Francis

[22] Black, P. and Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21 (1). 5-13. p9

[23] Adapted from Harlen, W. (2006) *Teaching, Learning and Assessing Science 5-12*. London: Sage. p 87.

[24] Leahy, S. and Wiliam D. (2012) From teachers to schools: scaling up professional development for formative assessment, in (ed) J. Gardner *Assessment and Learning*. London: Sage. 49-71, p 51.

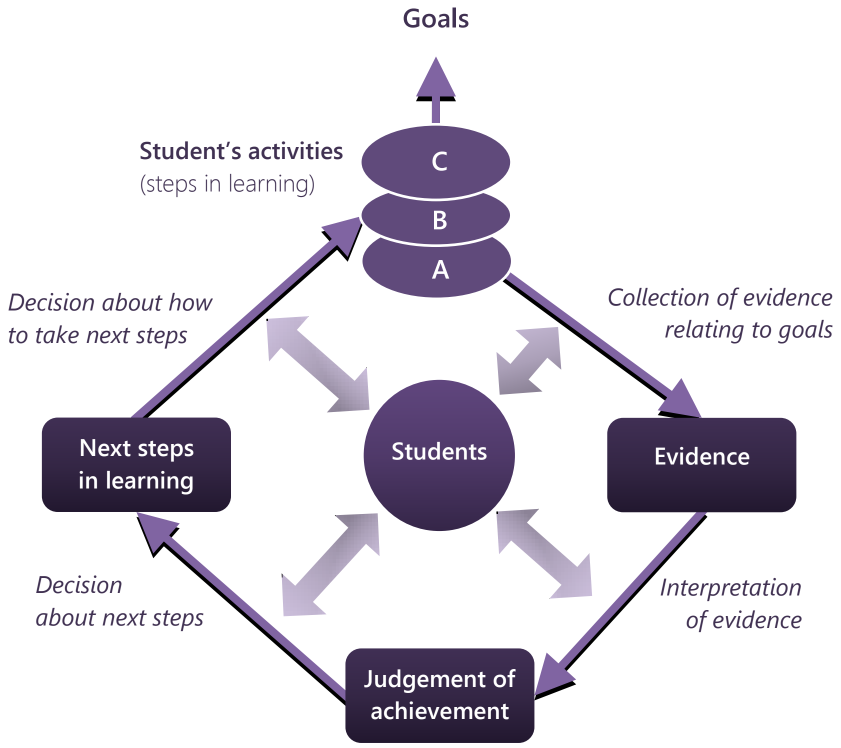


Figure 1:
***Assessment for formative purposes** (adapted from Harlen, 2006)*

The students are in the centre of the process, since it is they who do the learning. The two-headed arrows linking students to the various parts of the assessment cycle indicate that students both receive feedback from the teacher and also provide information. They participate in decisions where appropriate through self- and peer-assessment.

In formative assessment, judgements about progress and decisions about next steps take into account the circumstances, past learning and effort of individual students as well as what they are able to do in relation to the goals of the work at a particular time. Thus the judgements are both student-referenced and criterion-referenced (see Chapter 1). This approach supports learning far more than applying the same standards to all students, which would be demotivating for lower achieving students, and is possible since no comparisons are made between students in formative assessment.

The actions indicated by the arrows in Figure 1 are not 'stages' in a lesson nor necessarily the result of planned decisions made by the teacher. They represent the thinking involved in focusing on what and how students are learning and using this to help further learning. In some cases it may be possible for teacher and students together to decide on immediate action. In other cases, the teacher may take note of what help is needed and provide it at a later time.

Implementing formative assessment means that not everything in a lesson can be planned in advance. By definition, if students' existing ideas are to be taken into account, some decisions will depend on what these ideas are. Some ideas can be anticipated from teachers' experience and from research findings built into curriculum materials, but not all. What the teacher needs is not prescribed lesson content but a set of strategies to deploy according to what is found to be appropriate on particular occasions. Some illustrations of these strategies are given in Chapter 5.

Box 5: Key component practices of formative assessment

- Students being engaged in expressing and communicating their understandings and skills through classroom dialogue, initiated by open and person-centred questions
- Students understanding the goals of their work and having a grasp of what is good quality work
- Feedback to students that provides advice on how to improve or move forward and avoids making comparisons with other students
- Students being involved in self-assessment so that they take part in identifying what they need to do to improve or move forward
- Dialogue between teacher and students that encourages reflection on their learning
- Teachers using information about on-going learning to adjust teaching so that all students have opportunity to learn[25]

Feedback is an essential feature of formative assessment. The two-way feedback, from teacher to students and students to teacher, implies a view of learning as a process in which understanding is actively constructed by students:

- Feedback from teacher to students gives students information to help them take the necessary steps to improve their understanding or skills. The form and focus of the feedback has to be carefully judged by the teacher. The focus of the feedback influences what students pay attention to and the form it takes determines whether it can be used to advance learning. In a view of learning in which learning is equated with 'being taught' feedback to the student from the teacher is about the quality or success of the students' work rather than how to improve it. Formative assessment really has no role in learning seen this way.
- Feedback into teaching, from students to teachers, is necessary so that teachers can adjust the challenges they provide for students to be neither too demanding, making success out of reach, nor too simple to be engaging. Using feedback from observations of students and their work to judge the students' ability to take certain steps with help (the zone of potential development, as discussed in Chapter 4) is a complex and challenging task for teachers. Many teachers need a good deal of help with this task if they are to use feedback to regulate teaching in order to optimise learning.

In summary, the key components of formative assessment are summarised in Box 5. In Chapter 5 we consider how these components of formative assessment can be put into practice.

Why it is important

The importance of formative assessment lies in the evidence of its effectiveness in improving learning. Empirical studies of classroom assessment have been the subject of several research reviews. The review by Black and Wiliam (1998) attracted attention world-wide partly because of the attempt to quantify the impact of using formative assessment. Since then there have been a number of other reviews and investigations which have justified the considerable claims made by Leahy and Wiliam:

> The general finding is that across a range of different school subjects, in different countries, and for learners of different ages, the use of formative assessment appears to be associated with considerable improvements in the rate of learning. Estimating how big these gains might be is difficult… but it seems reasonable to conclude that use of formative assessment can increase the rate of student learning by some 50 to 100%.[26]

[25] Harlen, W. (2007) op cit p121

[26] Leahy, S. and Wiliam, D. (2012) op cit p52

Stobart,[27] however, strikes a note of caution, pointing out that, apart from a study by Wiliam *et al* (2004) of the impact of their action research project on student achievement, 'there is, as yet, little direct empirical evidence of the impact of [formative assessment] on achievement'. He notes that most evaluation studies have focused on the extent of change in teachers' practice and in students' attitudes and involvement rather than in students' conceptual learning. Nevertheless it can be argued that such changes are necessary steps towards improved learning. Moreover, the number of influences on students' measured learning, other than what may seem rather subtle changes in pedagogy when formative assessment is implemented, makes its impact difficult to detect. Indeed, Wiliam *et al* (2004) point out that the comparisons on which they base their claims are 'not equally robust'.

The importance for IBSE of formative assessment lies in the claim of both IBSE and formative assessment to support the development of real understanding and of competences needed for continued learning. Teaching for the development of understanding involves taking account of students' existing ideas and skills and promoting progression by adjusting challenge to match these starting ideas.[28] The practice of formative assessment, through teachers and students collecting data about learning as it takes place and feeding back information to regulate the teaching and learning process, is clearly aligned with the goals and practice of inquiry-based learning. It also supports student ownership of their learning through promoting self-assessment and participation in decisions about next steps, helping students to take some responsibility for their learning at school and beyond.

Summative assessment

Since we have just described formative assessment as having a positive role in learning, there is a tendency to consider it as the 'good ' face of assessment, with summative assessment, which has a different role, as the 'bad' face. This is unfortunate in several respects. First, whilst summative assessment is not intended to have direct impact on learning as it takes place, as does formative assessment, it nevertheless can be used to help learning in a less direct but necessary way as, for example, in providing a summary of students' learning to inform their next teacher when students move from one class or school to another. Second, it enables teachers, parents and schools to keep track of students' learning, both as individuals and as members of certain groups (such as those who are high achievers and those who need special help). Third, it provides data which, together with contextual factors, can be used for school evaluation and improvement. The bad reputation of summative assessment arises from inappropriate use of data which do not fully reflect the goals of learning. The danger is acute in the case of IBSE on account of its range of different goals, which are not easily assessed by conventional methods. This makes it even more important to consider the ways in which dependable information can be gathered for summative assessment.

What it is

Summative assessment is the name given to assessment that is carried out for the purpose of reporting achievement at a particular time. It may, and often does, have some impact on learning and the outcome may be used in teaching, but that is not its main rationale. We return to the relationship with formative assessment later, but for the moment represent the process as one of providing information purely for reporting achievement as in Figure 2.

[27] Stobart, G. (2008) op cit p154

[28] Bransford, J.D., Brown, A. and Cocking, R.R. (Eds) (2000) *How People Learn, Brain, Mind, Experience and S*chool. Washington, D.C.: National Academy Press.

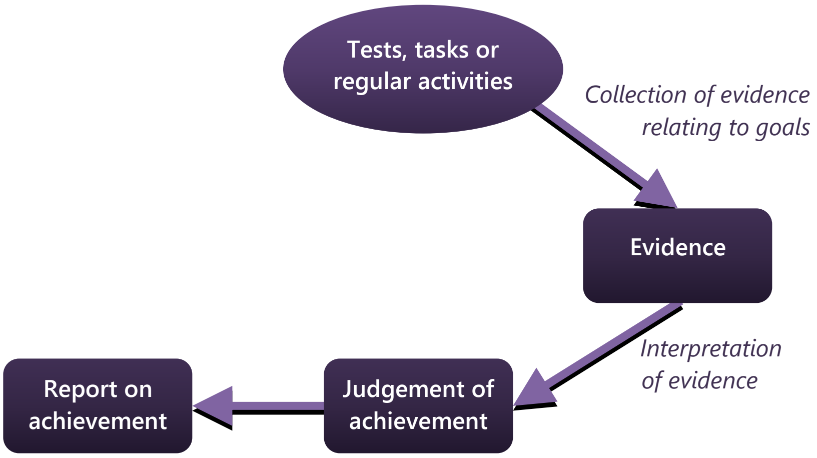

Tests, tasks or regular activities
Collection of evidence relating to goals
Evidence
Interpretation of evidence
Judgement of achievement
Report on achievement

Figure 2:
Assessment for summative purposes *(adapted from Harlen, 2006)*

The evidence derives from tests, special tasks or regular activities and can be collected by a range of means from different sources: written answers, artefacts constructed by students, portfolios, observation of actions, discussion or presentations of work. Clearly the collection of evidence about performance in relation to all relevant understanding and competences is the most important part of the process, for without it the final report on achievement is unlikely to provide dependable information about students' achievement of the goals of learning. The pros and cons of using different sources of data are discussed in Chapter 6.

The evidence is interpreted by comparison with criteria or standards relating to overall goals, rather than the goals relating to specific lessons or topics, as in the case of formative assessment. This marking or scoring can be carried out by the teacher or by an external agency, as in the case of some national tests and examinations. Only in the most informal classroom tests do students usually have a role in this process. Students are all judged by the same criteria, or mark schemes (rubrics), whereas, as noted earlier, in formative assessment criteria may be ipsative, or student-referenced (see Chapter 1) in order to help students recognise their progress from different starting points.

The interpretation necessarily reduces the richness of the actual performance to a score, category or mark that represents it; thus a great deal of information is lost. Depending on the use to be made of the result the process of interpretation will include some procedure for increasing reliability of the result. Where results are used to compare students, particularly where high stakes selection or grading is involved, steps are taken to check marking and moderate judgements by teachers or examiners. When the summative assessment is essentially classroom-based and in the hands of the teacher there is the potential for evidence to be collected and used about a wide range of kinds of achievement.

The absence of reference to students in Figure 2 acknowledges that generally they do not have a role in summative assessment. However, when the process is an open one and assessment criteria are shared with students and users of the results, and not restricted to what can be done in a test or controlled situation, there is an opportunity for students to have a role in the process, as for instance in selecting items in a portfolio. There remains, of course, the obligation to ensure that judgements are reliable and based on the same criteria for all students. We consider ways of doing this in Chapter 6.

Box 6: Key component practices of summative assessment

- Students may be involved in special tasks or tests as part of, or in addition to, regular work
- Takes place at certain times when achievement is to be reported, not a cycle taking place as a regular part of learning
- Relates to achievement of broad goals expressed in general terms rather than the goals of particular learning activities
- Involves the achievement of all students being judged against the same criteria or mark scheme
- Requires some measures to assure reliability
- Provides limited opportunities for student self-assessment

The form of report depends to a large extent on the nature of the task, the basis for judgement and the audience for the report. Numerical scores from tests are a summation over a diverse set of questions. The same total can be achieved in many ways, so scores have little meaning for what students actually know or can do. They also give a spurious impression of precision, which is very far from being the case. Scores can be used directly to rank order students, but this is really only useful in the context of selection since a position in a rank order gives no indication of meaning in terms of learning.

In theory, reporting against criteria which describe performance at progressive levels or grades can provide more meaningful indication of what students have achieved (see Chapter 6). However, In order to preserve some meaning in the report, a profile is preferable to a single overall grade or level which would have to combine different domains. The shorthand of 'levels' – labels given to progressive criteria - can be useful for some purposes, but for reporting to parents and students, the levels need to be explained or accompanied by accounts of what the student can do. Moreover, as noted later (Chapter 6, page 69) the use of levels can have negative implications for students' motivation and learning.

Some characteristic practices of summative assessment are summarised in Box 6.

Why it is important

Several reasons have already been given at the start of this section. The most compelling, however, is that summative assessment is important because it is necessary. It is unavoidable – reports on students' learning have to be made and records kept at regular intervals. By contrast, formative assessment could be considered, in a sense, to be voluntary, in that it is possible to teach without it. But teachers need to keep records summarising students' performance at key points, such as the end of topics or semesters, and to use these records in their planning. Parents and students' next teachers at points of transition from class to class or school to school need records of what has been achieved. School principals and managers need to have records so that they can review progress of groups of students as they pass through the school for use in school self-evaluation and curriculum planning.

It is also important because *de facto* what is assessed is taken as a signal of important learning. Unfortunately it is often the case that what is assessed is what *can be* assessed rather than what *ought to be* assessed. The gap between these two is likely to be particularly large in the case of IBSE where goals relate to building understanding and developing 'skills used by scientists'. We consider in Chapter 6 the 'pros and cons' of various methods of summative assessment but it is worth noting here that there is no 'perfect' approach or method. Any assessment is only a sample of what has been learned and an approximation of how well it has been learned and the result of several subjective judgements. A better understanding of the process of assessment by everyone involved – from teachers and students to politicians and employers – might help loosen the grip of assessment on the curriculum. We return to this in Chapter 4.

The relationship between formative and summative assessment

One reason for summative assessment gaining the reputation of being the bad face of assessment is that when measured performance becomes the dominant factor in the classroom it drives out formative assessment practice. The nature of the feedback given by teachers to students indicates more than the next steps that are needed but adds to the general impression that students have of their teachers' helpfulness and interest in them as learner. This is illustrated in the following report of research:

> Roderick and Engel[29] reported on how a school providing a high level of support was able to raise the effort and test performance of very low achieving and disaffected students to a far greater degree than a comparable school providing low level support for similar students. High support meant creating an environment of social and educational support, working hard to increase students' sense of self-efficacy, focusing on learning related goals, making goals explicit, using assessment to help students succeed and creating cognitive maps which made progress evident. They also displayed a strong sense of responsibility for their students. Low teacher support meant teachers not seeing the target grades as attainable, not translating the need to work harder into meaningful activities, not displaying recognition of change and motivation on the part of students, and not making personal connections with students in relation to goals as learning.[30]

Pollard et al[31] noted that the introduction of national tests in England and the requirement for teachers to assign levels to students affected their response to students and their use of formative assessment. Students were aware that whilst effort was encouraged, it was achievement on tests that counted. It is hardly surprising, then, that summative assessment has acquired a poor reputation, even though the problem often stems from the use made of the results rather than the nature of summative assessment itself. As we have noted, some summative assessment is necessary and unavoidable, leading to the question: are there ways in which summative assessment can be carried out and used without having a damaging impact on formative assessment?

An obvious approach is to think the other way around, that is, to focus on the formative assessment and see whether it is possible to derive data from this formative process to provide a summative judgement. Traditionally, the Danish educational system has worked this way. Not having any national testing in the compulsory school, teachers in the early grades provided feedback to students on their progress in an on-going process and in the higher grades the teacher summed up progress every four to six months as a mark. At the end of compulsory school the most recent mark appeared on the final certificate. For some subjects there was an oral examination with an external assessor, typically a teacher from a neighbouring school, and the mark from these examinations were also written on the final certificate. The obvious problems were, of course, how to secure reliability and students' protection against hostile teachers.

The two lists of characteristics of formative (Box 5) and summative assessment (Box 6) and the representations in Figures 1 and 2 show some key differences in the use made of evidence. However, as hinted earlier, there is the potential for evidence collected for a summative purpose to be fed back into teaching and learning to identify aspects where attention is needed to improve performance.

[29] Roderick, M. and Engel, M. (2001) The grasshopper and the ant: motivational responses of low achieving pupils to high stakes testing. *Educational Evaluation and Policy Analysis* 23: 197-228

[30] Harlen, W. (2012a) The role of assessment in developing motivation for learning, in (ed) J. Gardner *Assessment and Learning*, pp171-184. p177

[31] Pollard, A., Triggs, P., Broadfoot, P., Mcness, E. and Osborn, M. (2000) *What pupils say: changing policy and practice in primary education* (chapters 7 and 10). London: Continuum

Black et al[32] provide examples of teachers using classroom tests to enable students to identify their areas of weakness and focus further effort. In practice the approach is one that teachers can use principally in the context of classroom tests over which they have complete control. Whilst some external tests and examinations can be used in this way, by obtaining marked scripts and discussing them with students, there is a danger that the process can move from developing understanding to 'teaching to the test'.

An example of combining formative and summative purpose in assessment that has high stakes for students is the approach used for many years in the Senior Certificate in Queensland, Australia. It is important to recognise that this is designed to serve both purposes through building in participation of students and procedures for quality assurance. A portfolio of evidence, built up over the two years of the course, provides feedback to students enabling them to improve their performance during the course as well as showing what they have achieved by the end of the course. Maxwell explains that:

> For this approach to work, it is necessary to express the learning expectations in terms of common dimensions of learning (criteria). Then there can be discussion about whether the student is on-target with respect to the learning expectations and what needs to be done to improve performance on future assessment where the same dimensions appear.
>
> As the student builds up the portfolio of evidence of their performance, earlier assessment may be superseded by later assessment covering the same underlying dimensions of learning. The aim is to report 'where the student got to' in their learning journey, not where they started or where they were on the average across the whole course.[33]

The criteria for assessment are published so that students and parents as well as teachers can be familiar with them. They describe what students can do in various categories and sub-categories at five levels or standards (see example in Box 20 Chapter 6). Evidence from the portfolio is compared with the criteria using 'on-balance' judgements of best fit. Key conditions for such an approach are time for teachers to take part in moderation to ensure dependability of the results and respect for teachers' professionalism. It is also worth noting that success of students in the Senior Certificate is detached from school and teacher accountability procedures. Schools are nevertheless encouraged to use the data from the certification process for self-evaluation and school improvement.

These examples are of summative assessment providing information used formatively, which is one way of looking at the connection between assessment for these two purposes. A corollary is the use of evidence collected for formative assessment being used for summative purposes. Since formative assessment is in the hands of the teacher it follows that the evidence to be used in the summative assessment will be collected by the teacher, giving the opportunity to collect the wider range of evidence that classroom activities provide. An example of dual use of data collected by the teachers is given in Chapter 6 (page 68).

However, there are key differences that need to be taken into account in the way judgements are made when evidence is used formatively from judgements for summative assessment. Formative judgements are likely to be student-referenced as well as related to lesson goals, whereas judgements for summative assessment need to be only criterion-referenced using broader criteria relating to longer-term goals. It requires a distinction to be made between evidence and judgements, so that the *evidence* used in formative assessment is reviewed against broader criteria that define levels or

[32] Black, P., Harrison, C., Lee, C., Marshall, B. and Wiliam, D. (2003). *Assessment for Learning: Putting it into Practice*. Maidenhead, England: Open University Press.

[33] Maxwell, G. (2004) '*Progressive assessment for learning and certification: some lessons from school-based assessment in Queensland*.' Paper presents at the third conference of the Association of Commonwealth Examination and Assessment Boards, Redefining the Roles of Educational Assessment, March, Nidi, Fiji. p2-3

grades. We discuss the further steps needed to ensure dependable use of assessment by teachers for summative purposes in Chapter 6.

A final point to make about the relationship between formative and summative is to question whether there is any value in making a distinction between them or whether the relationship is better considered as a dimension rather than a dichotomy.[34] There are different ways of using formative assessment just as there are different ways of collecting evidence for summative assessment. Some formative assessment involves an immediate response to students' actions, whilst in other cases it requires some thought and planning. There is also a difference between whether the collection of evidence arises as part of the learning activity or is planned in order to find out what has been learned. Planned formative assessment might have some similarity to summative assessment by teachers. What would make it formative would be the use made of the information. Thus there appears some blurring of the distinction between formative and summative assessment and the relationship is perhaps better regarded as a dimension rather than a dichotomy. However, the importance of preserving the distinction lies in the role of assessment in helping learning for if we do not consider this then all assessment may become summative.

[34] Harlen, W. (2012b) On the relationship between assessment for formative and summative purposes, in (ed) J. Gardner *Assessment and Learning*. London: Sage 87-102. pp 97-100

Chapter 4
Assessment, pedagogy and the curriculum

One of the main reasons for this publication is the well-established relationships between assessment, the content of the curriculum and pedagogy (Figure 3). In the context of developing understanding and implementation of inquiry-based science education (IBSE) these relationships are of particular importance. Traditional methods and content of assessment rarely reflect the key goals of IBSE. Consequently assessment, as currently practised, tends to act as a brake on the implementation of inquiry-based education. The challenge here is to change assessment practices so that they have a supporting role in teaching and learning science through inquiry rather than a restraining one.

In the first section of this chapter we consider the potential positive and negative impacts of assessment on what is taught and how it is taught. Strong evidence of negative impact emerges from the use of assessment results for high stakes evaluation of teachers and schools. The objection to this use is based on the unfairness of using test results as the sole indicator of the quality of teaching. This is not to say that teachers and schools should not be held accountable for the achievements of their students, but other information needs to be taken into account, as discussed in the second section. The third part of the chapter addresses the issue of the view of learning on which traditional assessment is based, one which conflicts with the view of learning implicit in inquiry-based education. Finally we argue that assessment, well designed and implemented, can have a role in ensuring that education meets some of its major challenges in relation to global problems created by human activity.

Impacts of assessment

The relationships between assessment, pedagogy and content are frequently expressed as triangular, but often with 'curriculum' at one apex. If we take 'the curriculum' to mean all that is experienced by pupils at school then each of these is an aspect of the curriculum as in Figure 3. However, most current use of the term 'curriculum' implicitly refers to the content through separating it from pedagogy and assessment. Here we refer to 'curriculum content' to make clear that content is just one aspect of the curriculum experienced by students, the whole including content, pedagogy and assessment and the underlying assumptions about learning.

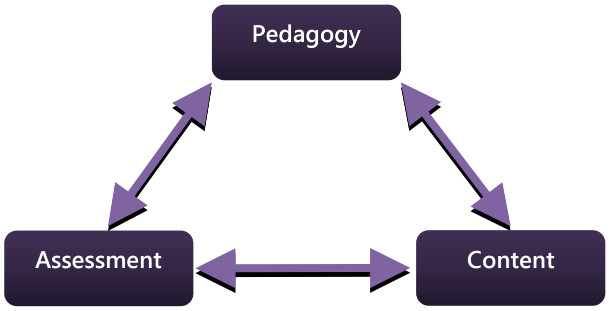


Figure 3:
Interactions among aspects of the whole curriculum

In Figure 3, the arrows acknowledge what is well known – that what we teach is influenced by how we teach, and what and how we assess influences both how and what we teach. These interactions are important, since it is no use advocating the use of inquiry-based teaching if there is an overbearing assessment (whether by testing or teachers' judgements) or a curriculum overcrowded with content. It is no use suggesting that the content should be focused on 'big' ideas if the assessment requires memorising multiple facts or if the pedagogy does not forge links that are necessary to form these big ideas; it is no use wanting pupils to develop responsibility for their own continued learning if teaching does not allow time for reflection and room for creativity. Nor can we hope for positive attitudes towards science if the curriculum content seems to pupils to be remote from their interests and experience.

The impact of assessment on the curriculum content and teaching approach is by no means necessarily a negative one. An effective assessment system supports learning in a variety of ways, from providing formative feedback for use in short-term decisions about learning activities to providing information about students' achievement for reporting to parents, for use in longer-term planning and as part of school self-evaluation. Through establishing criteria for achievement or providing tasks that exemplify the use of inquiry skills and understanding, assessment can help to clarify and communicate the meaning of learning objectives.

Negative impacts arise when what is assessed reflects only easily tested aspects of learning, compounded by applying rewards and punishments to the results. When test results are 'high stakes' for teachers this puts pressure on them, which is transferred to students, even if the tests are not high stakes for students. Research shows that when this happens, teachers focus teaching on the test content, train students in how to pass tests and feel impelled to adopt teaching styles which do not match what is needed to develop real understanding. There is now a large body of research evidence on the negative impact of high stakes use of data from assessment and testing. Box 7 gives a brief indication of findings from an extensive review of research on the impact of testing on teachers and on students.[35]

These findings raise questions about equity, since negative responses to tests are not spread evenly across all students. Students may be at a greater disadvantage on account of gender, language, home background and general ability.

A large-scale study of primary education in England, conducted between 2007 and 2009 and drawing evidence from a variety of sources concluded that the national tests students at the end of primary school (aged 11):

- put children and teachers under intolerable pressure
- are highly stressful
- constrain the curriculum
- subvert the goals of learning
- undermine children's self-esteem
- run counter to schools' commitment to a full and rounded education
- turn the final year of primary school into a year of cramming and testing.[36]

Similar effects are reported from other jurisdictions where the results of tests are used to make judgements on schools and teachers. However, research carried out in two states of Germany, Hesse and Bremen, suggests that curriculum narrowing and teaching to the tests may not result just from high stakes but accompanies the very existence of tests. The study by Jäger et al of the impact of low-stakes state-wide tests at the end of secondary education found that teachers were using feedback from the tests in an attempt to improve their students' results by reducing the curriculum breadth. The researchers suggest that 'This may be seen as an indication that teachers feel accountable for their students' performance and for reaching achievement goals and curriculum standards without external incentives.' [37]

[35] Nordenbo, S. E., Allerup, P., Andersen, H. L., Dolin, J., Korp, H., Larsen, M. S., et al. (2009). *Pædagogisk brug af test - Et systematisk review*. København: Aarhus Universitetsforlag. (In English: Pedagogical use of tests – A systematic review). http://www.dpu.dk/omdpu/danskclearinghouseforuddannelsesforskning/udgivelser/paedagogiskbrugaftest/

[36] Alexander, R. (Ed) (2010) *Children, their World, their Education*. Final report and recommendations of the Cambridge Primary Review. London: Routledge. P 316

[37] Jager, J.J., Merki, K.M., Oerke, B. and Holmeier, M. (2012) State-wide low-stakes tests and a teaching to the test effect? An analysis of teacher survey data from two German States, *Assessment in Education*, 19 (4) 451-467, p 464

Box 7: A systematic review of research on the impact of large-scale testing[35]

The Danish Clearinghouse for Educational Research performed a systematic review of research addressing the questions:

1. How does testing affect teaching?
2. How are tests used for pedagogical purposes by teachers?
3. How does testing affect the student?

The review was based on published research in the period 1980-2008 using a systematic methodology. The main findings revealed considerable negative effects on teaching of introducing centrally administrated tests:

- A narrowed down or distorted curriculum experienced by the students: teachers simplifying demands on students' thinking; facts and mechanical skills are emphasized at the expense of creative and aesthetic activities
- More teaching time being allocated to matters included in tests at the expense of those not included
- Teaching becoming devoted to teaching to the test and rote learning

The study also concluded that, in general, teachers do not use large scale tests or data from large scale tests in their teaching. Indeed they view such tests with skepticism and negative attitudes unless they have been involved in their formation. The very existence of large scale testing can make teachers less willing to make use of the test data. However, if teachers feel some ownership of the test they are more willing to use test data.

The influences of tests on students are dramatic:

- The mere announcement of a test starts emotional reactions such as nervousness and fear, especially among girls
- Students prepare for the test by learning by heart and memorizing sentences
- For high achievers motivation increases while low achievers lose their motivation
- A student's test result can influence future motivation and self-efficacy

The consequences for science are particularly serious. Teachers have been observed to adopt a transmission style of teaching even though this is not what they believe to be the best for helping students' understanding and development of skills.[38]

Another of the most serious impacts reported is on the practice of formative assessment. Following the introduction of a testing regime in England, researchers reported that students and teachers were aware that classroom assessment, rather than being essentially formative in function, became a series of mini-summative assessments.[39] Teachers checked the 'level' of performance far more often than was necessary for reporting. A 'performance oriented culture' developed, allowing little room for using assessment formatively. This culture was seen as responsible for schools voluntarily exceeding what was required in terms of external assessment and judgements about levels reached by students.

[38] Osborne, J., Simon, S. and Collins, S.(2003) Attitudes towards science: a review of the literature and its implications, *International Journal of Science Education*, 25, 1049-1079

[39] Pollard, A and Triggs, P. (2000) *Policy, Practice and Pupil Experience*. London: Continuum International

The influence that summative can have on formative assessment clearly means that giving attention to formative assessment alone would be likely to have little effect. Indeed the experience of the obstacles to introducing genuine formative assessment in countries where there exists a strong dependence on external high stakes tests, bears evidence to this. Thus if learning in science is to be improved through IBSE and the use of formative assessment, it is necessary also to ensure that the summative assessment is consistent with the learning aims of IBSE.

The myth of raising standards by testing

Given these negative effects, one may well ask: why do national tests have such a strong role in so many educational policies? The answer lies in the claim that 'testing drives up standards'. However, there is little evidence to support this and indeed plenty to suggest that rise in tests scores is due to familiarity with test-taking and to teaching to the test.[40] Test scores may rise – at least at first – but this does not give information about change in real learning. The consequence of focusing on what is tested, practising test-taking and the restricted range of what is tested, is that it is not really possible to tell from national test results whether or not national standards have changed year-on-year.

This phenomenon, where the impact of a measure becoming 'high stakes' is to occlude its value as an indicator of quality of a service, is a common one across several social services. It is summarised in a neat formulation of relationships as Goodhart's Law,[41] which is usually stated succinctly as

> When a measure becomes a target, it ceases to be a good measure.

In other words, once a social or economic indicator or other surrogate measure is made a target as part of policy it loses the ability to give valid information for its original purpose. There are several examples in other areas than education (for example in the health service and police services) where a single easily measurable outcome (such as waiting time for treatment in a hospital) is used to assess a complex process. In the case of education, tests are often the chosen means to set targets and the high stakes leads to practices that raise test scores without a genuine improvement in learning, thus the test ceases to be a useful measure of learning. Although most commonly illustrated in terms of tests, the same effect can be seen in assessment where performance is judged against criteria. High stakes use of the results leads to narrow interpretation of the criteria which then cease to indicate the full range of competences intended.

Accountability

These effects of judging teachers and schools on the basis of test results are now widely acknowledged and give rise to the claims that 'accountability undermines learning'. This is mistaken, however, since it is not accountability that is the problem but how it is exercised. When test scores are taken as the measure of quality of teaching this will inevitably have consequences for those being judged. Some of these consequences may be thought to be unintended, as when teachers interpret too narrowly what is tested as being important to teach. Those responsible for the accountability procedures may protest that this was not intended, that 'teachers do not need to teach to the test' and agree that some impacts of the test regime are undesirable. Yet the practice continues. One may well suspect that holding teachers accountable for raising test scores is a deliberate use of the impact of assessment on curriculum content; testing provides a direct way of controlling what is taught. As Stobart[42] comments, this is a quicker and cheaper way of changing curriculum content and pedagogy than developing curricula and providing professional development. The mechanism for this influence is to set targets for schools, districts or the nation as a whole to raise their test results.

[40] See, for example: Linn, R. L. (2000) Assessments and accountability, *Educational Researcher*, 29 (2) 4-16 and Tymms, P. (2004) Are standards rising in English primary schools? *British Educational Research Journal*, 30 (4) 477-94

[41] http://www.atm.damtp.cam.ac.uk/mcintyre/papers/LHCE/goodhart.html

[42] Stobart, G. (2008) op cit p138

At the system level the ambition of many countries to improve their position in the 'league table' of PISA results has led to actions which pass down to districts and schools the imperative to change practices.[43] At the national level, the use of tests appears to have the desired effect for a few years. The national tests for students at the end of primary school in England provide an example. Tests in English, mathematics and science for every 11 year old were introduced in 1995. For the first five years scores rose year by year, but from 2000 onwards there was no appreciable change. An exhaustive study[44] of other data about change in performance over these years (including data from TIMSS surveys) showed no rise in performance in these years. The conclusion reached was that the most likely reasons for the initial changes were the effect of teaching test technique (formal tests were new to pupils of this age in 1995) and teaching focused narrowly on tested content.

Similar finding and conclusions have been reported in other test regimes; for example in the USA Linn found 'a pattern of early gains followed by a levelling off'[45] to be typical across States where high stakes tests were used. Stobart describes the point of levelling off as the 'half-life' of accountability testing. This is probably around four years, during which:

> The slack has been taken up; teachers have become familiar with the test; and students know from past papers what they need to do, so learning is increasingly reduced to how to maximise marks.[46]

At the school level the requirement to meet targets in the form of test scores means that greater attention is given to those subjects tested than to others. For primary schools, where reading, writing and aspects of mathematics are invariably targeted for national testing, this has serious consequences for science and other subjects. Indeed, when end-of-primary school testing in science was terminated in England in 2010 – for reasons relating to the low validity of the tests, which were written and did not require any practical experience – the effect was the transfer of class time and attention to English and mathematics, which continued to be the subject of high stakes testing.

What is important is to avoid inappropriate use of student test results, such as when they are taken to be the sole indicators in the evaluation of teachers and schools. The reason for this is simply that what students achieve is not only the result of their experiences in school.

The limits of accountability

Being accountable means being responsible for one's actions and being able to explain to stakeholders why and how certain things were done or why they were not done. Teachers can only be held accountable for actions or outcomes over which they have control, such as what they do in the classroom, what learning opportunities they provide and the help they give to students, and so on. They are not necessarily responsible for whether externally prescribed learning outcomes are achieved, since this depends on many other factors, over which the teacher does not have control, such as the students' prior learning and the many out of school influences and conditions that affect their learning. These factors need to be taken into account both by teachers in setting and working towards their goals for students' learning, and by those who hold teachers accountable for the quality of students' education. It follows from these arguments that the information used in accountability should include, in addition to data on students' achievements, information about the curriculum and teaching methods and relevant aspects of students' backgrounds and of their learning histories.

[43] Ertl, H.(2006) Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education*, 32 (5) pp 619–634

[44] Tymms, P. (2004) Are standards rising in English primary schools? *British Educational Research Journal*, 30 (4):477-494.

[45] Lynn, R.L. (2000) Assessment and accountability, *Educational Researcher*, 29 (2), 4-16, p6

[46] Stobart,, G. (2008) op cit p134

Box 8: Schools giving their own account

A positive approach to accountability enables schools and teachers to express their own goals and their achievements in relation to them. Schools are now expected to communicate and explain their philosophy, aims and policies to the wider community. This is an opportunity for information about the achievement of students to be presented within the context of the many factors that affect this achievement. The process is facilitated in some countries by guidelines[47] for schools giving a framework for setting out their own account, designed to foster schools' self-evaluation and a more formative role for the whole process. In some cases this is checked by inspectors; in others it is available for public scrutiny.

Data on student achievement, derived from teachers' records and assessment relating to a wide range of goals of learning, not only test scores, will be part of a school's account of its performance to parents and others with an interest in the school's performance. The account should also show how, as part of their internal school evaluation, aggregated data about the achievements of students is used for checking the progress of groups of students such as those with different home languages, those who are higher and lower attaining, and boys and girls. All this depends on having effective measures in relation to the learning goals and setting the results in the context of the actions and systems used to work towards these goals.

Assessment and theories of learning

The alignment of assessment with curriculum content and pedagogy is important in claims for validity of the assessment. The discussion in Chapter 1 indicated that validity is a judgement of the extent to which inferences drawn are supported by evidence and theoretical rationales. A point expressed frequently in the conference (see Introduction) was the need for assessment to reflect the theory of learning which underpins IBSE. We need, then, to bring learning theories into the discussion of the impact of assessment on curriculum content and pedagogy.

Learning theories

The many and various theories of learning can be grouped into three kinds: 'behaviourist', 'cognitive constructivist' and 'socio-cultural constructivist'. (In some US literature[48] constructivist is described as 'cognitive' and socio-cultural as 'situated', but the meanings are largely the same.) A simple formulation of these terms, based on Watkins,[49] expresses their meaning as:

- Behaviourism: "Learning is being taught"
- Cognitive constructivism: "Learning is individual sense-making"
- Socio-cultural constructivism: "Learning is building knowledge as part of doing things with others."

Behaviourism describes a view of learning in which behaviours are formed by a system of rewards and punishments. The assumption is that behaviours that are rewarded will be reinforced and those that are punished will disappear; learning can be controlled externally and motivation is almost entirely extrinsic. A further feature of behaviourism that is particularly relevant to assessment is that complex behaviours are deconstructed into parts which can be taught, practised and assessed separately. This view, then, is consistent with tests of disconnected facts and skills, where speed is of the essence and answers are either correct or incorrect.

[47] SEED (Scottish Executive Education Department) (2002) *How Good is Our School? Self evaluation using quality indicators*. Edinburgh: HMIE.

[48] Discussed in Pellegrino, J.W., Chudowsky, N. and Glaser, R. (Eds) (2001) *Knowing what Students Know - The Science and Design and Educational Assessment*. Washington, DC: National Academy Press.

[49] Watkins, C. (2003) *Learning: A Sense-Maker's Guide*. London: Association of Teachers and Lecturers.

Two key features of *cognitive constructivist* views of learning are that learners construct their own understanding by developing mental models and that existing knowledge has an important role in this development. The aim is understanding, which is seen as occurring when new experience is incorporated into an existing or new model. The active participation of students is seen as paramount because, as widely quoted, 'they do the learning'. Constructivist views of learning underpin formative assessment, which starts from 'finding out where learners are in their learning' in order to decide 'where they need to go and how best to get there' (see page 18). There are few examples of summative assessment being based on a constructivist view of learning, although there are some attempts through computer adaptive testing and screen-based concept-mapping.[50] James concludes that 'much formal testing still relies heavily on behaviourist approaches'.[51]

In *socio-cultural constructivist* perspectives on learning there is also a focus on understanding but through 'making sense of new experience with others' rather than by working individually. In these situations the individual takes from (internalises) a shared experience what is needed to help his or her understanding, then externalises the result as an input into the group discussion. There is a constant to-ing and fro-ing from individual to group as knowledge is constructed communally through social interaction and dialogue. Physical resources and language also have important roles, as James explains:

> According to this perspective, learning occurs in interactions between the individual and the social environment. Thinking is conducted through actions that alter the situation and the situation changes the thinking; the two constantly interact. Especially important is the notion that learning is a *mediated activity* in which cultural artefacts have a crucial role. These can be physical artefacts such as books and equipment but they can also be symbolic tools such as language. Since language, which is central to our capacity to think, is developed in relationships between people, social relationships are necessary for, and precede, learning (Vygotsky, 1978). Thus learning is a social and collaborative activity in which people develop their thinking together.[52]

Some profound implications for assessment follow from Vygotsky's[53] view that for any learner there is an area just beyond current understanding (where a person is in conscious control of ideas and knows that he or she is using them) where more advanced ideas can be used with help. Vygotsky called this area the 'zone of proximal (or potential) development'. It is, in essence, what we have called the 'next step' that the student can be expected to take identified through formative assessment. 'Scaffolding' is an apt term used to describe helping students to take this next step in understanding through introducing new ideas or better scientific practices and providing vocabulary that enables students to express their ideas more precisely.

Recognising that, in the company of other learners, students can exceed what they can understand and do alone, throws into doubt what is their 'true' level of performance. Is it the level of 'independent performance' or the level of 'assisted performance' in the social context? It has been argued that the level of performance when responding to assistance and the new tools provided by others gives a better assessment than administering tests of unassisted performance.[54]

[50] Osmundson, E., Chung, G., Herl, H., Klein D. (1999) Knowledge-mapping in the classroom: a tool for examining the development of students' conceptual understandings. Los Angeles, California: National Centre for Research on Evaluation and Student Testing, University of California. www.cse.ucla.edu/Reports/TECH507.pdf

[51] James, M. (2012) Assessment in harmony with our understanding of learning: problems and possibilities, in Ed J.Gardner *Assessment and Learning*, 2nd edn. London: Sage pp187 – 205.

[52] ibid p 192/3

[53] Vygotsky, L.S. (1978) *Mind in Society: The Development of Higher Psychological Process*. Cambridge, MA: Harvard University Process

[54] Grigorenko, E. (1998) Mastering tools of the mind in school (trying out Vygotsky's ideas in classrooms), in Eds R. Sternberg and W. Wiliams *Intelligence, Instruction and Assessment: Theory and Practice*. Mahwah, NJ: Erlbaum.

Box 9: A model of learning science through inquiry

In inquiry-based learning the development of understanding stems from curiosity about a phenomenon or event that is new to the learners and raises questions that grab their attention. Initial exploration may reveal features that bring to mind an idea from previous experience which suggests a possible explanation or an answer to a question. It may be the idea of an individual student or the result of brain-storming with other students or consulting sources of information. Working scientifically involves making a prediction based on the idea and then gathering relevant data to see if there is evidence to support the prediction and the application of the idea. This might be a lengthy investigation involving controlled experimentation or just a simple extension of observations.

Finding that evidence fits with the prediction and that the idea does provide a good explanation means that this idea has become 'bigger' since it then explains a wider range of phenomena. Even if it does not seem to 'work', something has been learned about its range of application. But to find an explanation that does 'work' means that alternative ideas have to be used and tested. This may come from the initial or further brain-storming informed by what has then been found. The usefulness of the ideas developed in this way depends on the collection and use of evidence in a scientific manner. Thus the ability to use science inquiry skills is an essential part of the development of understanding and an outcome of shared thinking about what data to collect and how to go about collecting and interpreting them.

Implications for assessment of IBSE

In Box 9 IBSE is described as a process of building understanding through collecting evidence to test possible explanations and the ideas behind them, which has elements of both constructivist and socio-cultural views of learning. If validity depends on how well assessment reflects the view of learning implicit in IBSE, how ought summative assessment of inquiry-based learning in science to be carried out? Evidently it is not by students sitting in isolation from each other in an examination room. Assessment consistent with the view of learning underpinning IBSE would be based on what students can do in interaction with others. When classroom or laboratory activities involve scaffolding by teachers and interactions among students, such activities are equally opportunities for assessing students' 'assisted performance' and add a further argument for teachers having a central role in assessment.

An example of assessment informed by social interaction and use of artefacts is the approach used by Dolin and Krogh in their research involving the re-examination of students on some PISA items (see page 60). This included an interview/conversation conducted with each student lasting about 30 minutes, exploring students' knowledge and reasoning. Such extended one-to-one situations are clearly of limited application where large numbers of students are assessed, however, and there are problems in respect of reliability. A more controlled situation would be for pairs or groups of students to be given a task and some time for brain-storming ideas about how to go about it and, if feasible, for conducting an inquiry. Each student then produces an individual account of the work which is assessed using a range of criteria.

A different approach might be for the 'interaction' to take place at a distance. Students conduct an investigation and receive comments on it which are returned to the student. The students' responses to the comments contribute to the overall assessment. Such approaches, of which there are few examples in practice, challenge conventional views of both the process and the meaning of assessment. As James notes 'There is still much work to be done to find ways of bringing assessment into better alignment with some of the most powerful ideas in contemporary learning theory'.[55] What may be needed for change to take place is a matter taken up in Chapter 7.

[55] James, M. (2012) op cit p196

The role of assessment in understanding major global problems

It seems to be stretching the impact of assessment too far to claim that it can help in developing understanding of major global problems, such as global warming, loss of diversity of organisms, starvation and poverty caused by human activity. Yet education surely has a key role in making changes to address these problems. Thus assessment, having a key role in education, must also have a part.

A crucial goal of education is to prepare students to be informed citizens who understand the reasons for problems created by human activity and what is required to solve them and are motivated to participate in responsible action. There are many aspects to consider but two important questions are:

- What are the essential ideas or concepts that need to be understood?
- How is understanding (as opposed to superficial knowledge of related facts) of these essential or core concepts to be engendered?

To answer the first question we need to identify the ideas that are relevant and powerful in helping understanding of the world and how it works, how its components interact, how human intervention can and cannot influence our global environment. This means identifying the 'big' ideas *of* science and *about* science (that is, how science operates, its strengths and limitations) and ensuring that science education is designed to develop understanding of these ideas.

The best answer we have to the second question, of how to pursue real understanding, is for students to learn through inquiry. We have described in Chapter 2 and in Box 9 the meaning of learning through inquiry, arguing that it enables active learning (physically and mentally) through which learners make sense of their investigations of the world around. This does not mean learning by themselves, or having to work out everything for themselves - far from it for, as noted earlier, current views of how ideas are developed point to the role of collaboration, discussion and dialogue with others, in which ideas are advanced by collective thinking and by building on what has already been found from the activities of scientists over the generations. What it does mean is that individual learners are developing shared understanding that makes more sense to them than alternative views that they (or others) may previously have held.

So where does assessment come into this argument? In brief, if the principle that 'all assessment should ultimately help learning' (see Box 21 in Chapter 7) is translated into practice, its role becomes clear. In more detail, using assessment formatively regulates teaching and learning so that understanding is supported. It acts to optimise the challenge of students' new experiences so that they are neither too distant from their existence ideas and competences nor too familiar so that there is little change in what they can do or know. In relation to summative assessment there are many ways in which this can be used to support learning. For instance, as mentioned in Chapter 3, mapping the progress of individual students and groups of students alerts all involved as to the need for action to ensure learning opportunities for all, for girls and boys, for the most able and the least able, for the disadvantaged as well as the more fortunate. What is mapped should reflect the aim of ensuring that all students are making progress towards the powerful ideas and the scientific inquiry skills that are part of the rationale for promoting IBSE. (Later, in Chapter 6, we note some implications of embracing this aim for the form of reporting students' achievements). The progress of students towards these goals is a key factor in schools' self-evaluation of their provision for students to learn with understanding.

Assessment is but one of a range of factors that influence students' learning. Clearly what students can achieve is dependent on the curriculum content and pedagogy, but the interactions indicated in Figure 3 and the earlier discussion in this chapter show that assessment cannot be ignored. Moreover it has a contribution to make to better understanding of the goals through expressing what it means to achieve them. Assessment, then, needs to be part of the discussion of how to provide education of relevance to facing global problems.

Chapter 5
Implementing formative assessment of IBSE

Chapter 3 considered the meaning and importance of assessment *for* learning (formative) and *of* learning (summative) in general terms. In this chapter and the next we look at methods of conducting assessment for these purposes but specifically in the context of assessing the goals of IBSE. In Chapter 2 six goals of science education were identified as essential for preparing students for life in the rapidly changing world today, whilst the definition of IBSE on page 12 highlights its specific contribution to the development of *understanding* and of the *skills used by scientists* (which we will call *science inquiry skills*). Thus the focus in this chapter is on how assessment can support the development towards these two goals: scientific understanding and science inquiry skills. Chapter 6 focuses on methods of summative assessment and reporting progress towards these goals.

Formative assessment and IBSE

Implementing formative assessment is not easy. It is not a matter of using particular materials or techniques, but requires skills and knowledge on the part of the teacher. The purpose of formative assessment is to identify, and to help students to take, next steps in progress of developing their understanding and competence. To implement formative assessment teachers require knowledge of how to gather and use evidence about students' progress in learning and how to provide effective learning environments that support further progress. Some researchers, following Sadler,[56] describe this as taking action to enable the student to 'close the gap' between the present state of understanding and competence and the learning goal. However, referring to 'next steps' rather than 'closing a gap' is a better way of conveying a view of progress in learning as a continuing process. In addition, and crucially, formative assessment requires a change in how teachers see the process of learning and their role within it. When learning is seen as something that *students do*, not something that is *done to* them, the teacher's role is to design environments in which students can be actively engaged in constructing their understanding and developing competences.

It is easy to see that formative assessment is essential to the implementation of IBSE. Learning through inquiry is a process of developing understanding which takes account of the way in which students learn best, that is, through their own physical and mental activity. It is based on recognition that ideas, knowledge and understanding are constructed by students through their own thinking about their experiences. What is known about learning tells us that this happens when students' activities enable them to develop their understanding, that is, when they are working in the area between existing and more advanced ideas and competence, or the zone of potential development (see page 32). Formative assessment is the strategy by which these activities and the whole learning environment are designed to ensure that this progress in learning is possible. When learning through inquiry students develop their understanding through the activities indicated in Box 3 and in the definition of IBSE: making observations, raising investigable questions, planning and conducting investigations, reviewing evidence in the light of what is already known, drawing conclusions and communicating and discussion with others in which ideas are shared, explained and defended. It is also recognised that engagement in these activities depends on the extent to which they hold interest, have perceived relevance and provide enjoyment and even excitement, for students.

[56] Sadler, D. R. (1989) Formative assessment and the design of instructional systems, *Instructional Science*, 18, 119-44.

For the teacher, formative assessment in IBSE means using the strategies identified in the description in Chapter 3:

- promoting classroom dialogue
- using questioning to generate evidence of, and to help the development of, students' ideas and competences
- providing feedback to students
- using feedback from students to regulate teaching
- encouraging students to participate in assessing the quality of their work.

We consider here how these aspects of formative assessment can be implemented by teachers. Later, in Chapter 7, we return to the matter of how to bring about the changes often needed in teaching in order to implement these practices.

Classroom dialogue

It is through language that we develop a shared understanding of ideas. The ideas that we may form from direct experience have to be communicated and this involves trying to find words that convey our meaning to others. In this process, our own ideas often have to be reformulated in ways that are influenced by the meaning that others give to words.

The value for learning of talking was established in ancient times, but in modern times it was Douglas Barnes in the 1970s whose research drew attention to the importance of informal or 'exploratory' talk.[57] In this kind of talk students interrupt each other, repeat themselves, hesitate and rephrase. Barnes suggested that students only engage in this kind of talk in the absence of the teacher because then there is no source of authority to which pupils can turn. However, he also showed that the teacher has a role in ensuring that thinking is probed, vagueness challenged and that students have a basis for their claims. The teacher, then, has to tread a careful path between domination of the discussion that would discourage students from developing and expressing their own ideas and leaving students without the stimulus to extend their thinking.

The extract in Box 10 is from the transcript of the discussion by a group of three girls observed by Barnes, before and during intervention by the teacher. Barnes points out that in asking the questions 'which air pressure?' 'why didn't it come up before?' 'at what point did it stop?' the teacher was persuading the girls to consider the process systematically, by providing the questions they had not asked themselves. However, the exchanges initiated by the teacher illustrate a common pattern in classroom discourse, in which:

- the teacher asks a question (which air pressure?)
- the student responds (inside the bottle)
- the teacher makes an evaluative comment (All right….),

then

- asks another question (why didn't it come up before then?)

and the pattern of 'question-response-evaluation' is repeated. This discourse involves plenty of interaction between teacher and students but in it the authority of the teacher is unmistakable. Contrast this with the role of the teacher in Box 11 where the teacher also encourages deeper thinking, use of evidence and clarity of meaning, but by sharing in the exchanges rather than dominating them. The teacher does not evaluate the students' responses but prompts them to express and explain them.

[57] Barnes, D. (1976) *From Communication to Curriculum*. Harmondsworth: Penguin.

Box 10: Explaining an effect of air pressure

After some lessons on air, the students (aged 12 to 13) were asked to carry out some simple activities that involved air pressure. In one of these the equipment was a bottle containing some water, closed by a stopper with a straw through it ending just below the surface of the water in the bottle. The students (T, B and C) were asked to blow strongly into the straw and then stand back, then to discuss and try to explain what happened.

Discussion without the teacher

T Ugh! The water comes up the straw.
C I wonder how that … I wonder why … How do you stop it?
? …it'll stop

Another girl blows into the straw

B Pr… probably it's 'cos of the air pressure
T Yes … it'll be the air pressure and …
C … it's the air pressure
C It's pressing down on the water.

Discussion when the teacher joins the group

B *(in answer to the teacher asking for what happened)*
T blew down it and it bubbled up and then she took her mouth away and it all came up because of the air pressure

Teacher Which air pressure?

B The … er … inside the bottle

Teacher All right. Now why didn't it come up before then?

C and T 'cos there wasn't enough air … air pressure

Teacher … There wasn't enough air in, but when you blew into it there was more air and it forced … Why did it stop? It's not going now … why isn't it going now?

Alexander has identified a role for the teacher in such verbal interactions, in the form of 'dialogic teaching'. He describes this as 'a distinct pedagogical approach', which 'harnesses the power of talk to stimulate and extend children's thinking, and to advance their learning and understanding. It also enables the teacher more precisely to diagnose and assess.'[58] It is through dialogic teaching that teachers can 'steer classroom talk with specific educational goals in mind'.[59] In relation to science this 'steer' focuses on the use of evidence and may lead to what has been described as 'argumentation'. This is different from argument in daily life:

> In science, goals of argumentation are to promote as much understanding of a situation as possible and to persuade colleagues of the validity of a specific idea. Rather than trying to win an argument, as people often do in non-science contexts, scientific argumentation is ideally about sharing, processing and learning about ideas.[60]

[58] Alexander, R. (2004) *Towards Dialogic Teaching. Rethinking Classroom Talk.* Cambridge: Dialogos, p1

[59] ibid: p 27

[60] Michaels, S., Shouse, A.W. and Schweingruber, H.A (2008) *Ready, Set, Science! Putting research to work in K-8 Science Classrooms*, Washington: National Academies Press, p89.

Box 11: Dialogic talk about eggs[61]

Two ten year old girls, D and M, were investigating the way in which three whole hens' eggs, labelled A, B and C behaved in tap water and in salty water. They knew that one was hard-boiled, one soft-boiled and one raw. They had to find out which was which.

When the eggs were placed in the salty water, the one labelled B was floating at the surface, C was at the bottom and A suspended about halfway down. The transcript begins with the teacher approaching them after they had been working alone for some time.

Teacher Can you tell me how you're getting on?
D I think that C is raw
M We both think that C is raw
Teacher Do you?
D B is ...
Teacher *(to D)* Why do you think that?
M Because when you put eggs in water bad ones rise to the top
D *(at the same time)* Because it ... we put them all in ...
Teacher Bad?
M Yes, I think so – or it is the good ones? ... well, I don't know
Teacher Yes?
M ... they rose to the top, so...
(D puts the eggs into the salty water)
D ... that's the bottom *(pointing to C)*
M ... if it's raw it should stay at the bottom
Teacher I see
D So that's what we think, C is raw and B is medium and A is hard-boiled
M ... and I think that B is hard-boiled and she thinks that B is medium
Teacher Ah, I see *(to D)* Can you explain, then, why you think that?
D If we put ... er ... take C out *(takes C out, puts it on the table, then lifts A and B out)* and put these in, one after the other. Put A in – no B first. That's what ... Allyson thinks is hard-boiled, I think it's medium. If you put that in ... *(she puts B into the salty water)*
M ...'cos it comes up quicker
D It comes up quick. And if you put that in...
(She puts A into the salty water. It goes to the bottom and rises very slowly)
M And that one comes up slower
D So, I think that one *(pointing to A)* is hard-boiled because it's ... well ...
M I don't. I think if we work on the principle of that one *(pointing to B)*. Then that one comes up quicker because it's, you know, not really boiled. It's like a bit raw
Teacher A little bit raw
M So, therefore, it'll come up quicker
D Yes, but it's not bad
Teacher What'll it be like inside?
M Runny
Teacher It'll be runny still, I see

[61] Quoted from Harlen, W. and Qualter, A. (2009) *The Teaching of Science in Primary Schools.* 5th edn London: Routledge, p 100-101

Box 11 gives an example of dialogic talk among a teacher and two girls (aged 10) trying to use evidence to distinguish between raw and cooked eggs. Having agreed that C is the raw egg, D and M disagree about the identity of the other two eggs. M has a reason for considering B is hard-boiled on the basis that 'bad ones rise to the top', so she considers that egg B behaves as if it had had something done to it. But she does not articulate the consequences of this until D attempts to give her reason. Then it is as if D's reason, which she interrupts, sparks off her own thinking.

The teacher does little here except to encourage the girls in their struggle to work out their answer and to explain their reasoning. Just the occasional 'Why do you think that?' the acknowledgement 'I see', and reinforcement 'A little raw', encourages their use of evidence in support of their arguments. This comes through most clearly in M's 'if we work on the principle that…' where she relates what she predicts on the basis of her judgement to the observation of how quickly the egg floats up in the salty water, but it also occurs throughout. It is worth noting in passing that the origin of her idea is previous knowledge about how to distinguish 'good' from 'bad' eggs.

In formative assessment, where the aim is for the students to reveal their developing understanding of a phenomenon or event, students' talk is a key source of information. The kind of exploratory thinking and dialogue which Barnes and Alexander have advocated is encouraged in a classroom climate in which teachers:

- expect students to explain things
- value their students' ideas even if these are unformed and highly conjectural
- avoid giving an impression that only the 'right' answer is acceptable and that students should be making a guess at it
- judge when to intervene or when it is better to leave discussion among students to proceed without interruption.

Teachers' questions

Questions have a central role in classroom discourse, both questions asked by the teacher and those asked by students of each other and the teacher. Questioning takes up a high proportion of teachers' talk and is one of the most important factors in determining students' opportunities for developing understanding through inquiry. It is not the frequency of questions that matters, but their form and content and how they feature in the patterns of classroom discourse.

In relation to *form*, the most relevant distinctions are between 'open' and 'closed' questions and between 'subject-centred' and 'person-centred' questions. Open questions allow students to express their view or observation ("what do you notice about …?") rather than responding to a particular point raised by the teacher ("are these…all the same size?"). Subject-centred questions ask directly about the subject matter ("why does this… take more time than …?"), whilst person-centred question ask for the student's ideas ("why do you think this…takes more time than…?"). The open and person-centred questions are more likely to give the teacher information about what students are noticing and thinking that may well be important in deciding how to help them.

In relation to *content*, questions need to be matched to the purpose of asking. We should not ask questions without a reason and without interest in the answer. If the answer is to be useful then it has to give the kind of information or stimulate the kind of response required. Questions are used in all parts of the formative assessment cycle (Figure 1 page 18) but, in the context of IBSE, most particularly in revealing and helping students to take next steps in developing their ideas and skills and in encouraging collaboration, sharing ideas, reflection and evaluation. The content will be different according to the focus at a particular time. For example questions such as:

What do you think is the reason for ...?

Why do you think this happens?

ask directly for students' ideas about how things can be explained, whilst

What do you see happening here?

What do you think will happen if....?

can be answered without giving an explanation.

Questions relating to inquiry skills will be worded according to relevant skills in a particular situation. A question designed to encourage students to make a prediction might be 'what do you think will make this plant grow faster?', whilst to encourage interpretation of data it might be 'what do you think made a difference to how fast this plant grew?' Such questions, which encourage use of science inquiry skills, have to be distinguished from those asking about the ideas developed from the inquiry or those designed to encourage application of ideas to other situations than those investigated.

Some examples of questions for different purposes, using as a context some activities students investigating shadows, are given in Boxes 12, 13 and 14

To gain access to students' ideas a good way to start is by using open questions. If this leads only to a vague response it may be necessary to follow up with rather more focused, but person-centred questions. For helping student to take next steps useful questions are likely to follow up the ideas they have expressed and be designed to encourage students to link different observations to see if their ideas work in other circumstances. They may also encourage students to consider different ideas, by 'scaffolding' the use of alternatives. These questions are likely to be rather less open than questions for finding out ideas.

Box 12: Questions for finding out and encouraging the development of ideas

Finding out

- What do you think makes the shadow?
- Why do you think these things make darker shadows than these?
- How do you explain the shape of the shadow?

Helping next steps

- How would your idea explain why the colour of the shadow is the same for all the objects?
- What other idea could explain the shape of the shadow?
- If the object cuts off the light from the wall how would this explain why the shadow is bigger if the object is closer to the torch?

The response to questions for finding out about students' inquiry skills, in Box 13, may be actions as well as words, giving the teacher opportunity to see what skills the students already have. The questions suggested for finding out also provide opportunity for development of inquiry skills. Again, they are open and phrased to encourage the students to answer without seeking for a 'right' answer. Some may involve scaffolding, providing support for thinking along certain lines, such as about variables or relating what they found to their initial question.

Box 13: Questions for finding out and encouraging the development of inquiry skills

Finding out

- What would you like to find out about shadows? (raising questions)
- What do you think will happen if we move the object this way? (prediction)
- What could you do to find out what makes a difference to the size of the shadow? (investigation)
- What have you found out about whether there is a connection between the position of the torch and the size of the shadow? (interpretation)

Helping next steps

- What do you find if you measure the size of the shadow before and after moving the torch?
- If the shadow gets bigger when you move the torch this way, what do you think will happen if you move it the other way?
- How will you make sure that it is the position of the torch and not something else that makes the difference?
- What have you found out about how you can change the size of the shadow?

Questions in Box 14 are designed to encourage students to work in genuine co-operation not as individuals – even though working groups or pairs – or in competition with each other. They require a response that reflects combined thinking, which might be about how to explain some observations or how to plan an investigation, etc. Questions for reflection and evaluation aim to ensure that students replay in their minds what they have done and so become conscious of how their ideas have changed. Without this reflection their ideas are likely to slip back to their previous way of thinking. These questions require children to talk about what they have learned and how they have learned so that they 'learn about learning' as well as about the things they have investigated.

Box 14: Questions for encouraging collaboration, sharing ideas, reflection and evaluation

Collaboration and sharing ideas

- How many different ideas can your group suggest to explain what you found?
- After using different ideas to explain what you found, is there one that seems best?
- Of all the ideas, which ones could you test?
- What do you agree is the way to find out which idea works?
- What would each person in the group do in this investigation?

Reflection and evaluation

- What have you found out that you did not know before?
- Have you changed your mind about …?
- What made you change your mind?
- What is there that you don't understand about … ?
- Is there something that you still want to find out?
- If you did this again what would you change so that you could learn more?

Time for answering

Carefully worded questions deserve thoughtful answers and students need to be allowed time to respond to a question. The questions suggested in Boxes 12, 13 and 14 are designed to provoke thinking; they require a thoughtful response. There is, perhaps, a place for the quick-fire test of memory, or quiz, but that is not what we are concerned with in the context of learning through inquiry. Pressure to respond quickly reduces the value of questioning for the purposes we have been discussing. So it is necessary to signal to children that a thoughtful, not a quick, response is required. This can be done in several ways.

- The first is to increase the 'wait time', the time between asking a questions and receiving a response. Teachers often expect an answer too quickly and in doing so deter students from thinking. The well-known research by Budd-Rowe[62] showed that extending the time that a teacher waits for students to answer has a marked effect on the quality of the answers. She found that teachers waited on average less than one second after asking a question, if no response was offered, before rephrasing or giving a hint or asking an easier question. When the teachers were asked to increase the wait time to eight or nine seconds, the quality of students' responses increased dramatically.
- The second is to avoid rephrasing a question if it is not readily answered. Putting the question a different way inevitably makes it more closed and less useful. When this happens regularly students realise that, if they wait, the teacher will ask a simpler question, often indicating the answer expected.
- Some teachers find it best not to allow students to raise their hands to answer these kinds of thoughtful question. They expect everyone to be able to answer, given time to think. So the teacher allows thinking time and then calls on students by name to contribute. This signals that thought, not speed, is valued and encourages all students to give attention to the question and so become more engaged in hearing other students' answers.
- Another strategy, suitable for some questions and situations, is to suggest that students discuss their answers with a partner or group for two or three minutes before the teacher asks for contributions.

Feedback to students

Feedback has been described as 'one of the most powerful influence on learning and achievement' but with the added warning that 'this impact can be either positive or negative'.[63] It has a key role in formative assessment since it is the mechanism by which future learning opportunities are affected by previous learning. Feedback is most obviously given by teachers to students orally or in writing, but also, perhaps unconsciously, by gesture, intonation and indeed by action, as when assigning tasks to students. How teachers provide feedback and what it focuses on is influenced by their view of learning (see Chapter 4). Constructivist view of learning lead to interaction between teacher and students in which students respond to the teachers' comments and suggestions rather than the one-sided communication from teacher to student that is typical of a behaviourist view of learning.

There are two main aspects of feedback to students to consider – the form it takes and the content.

[62] Budd-Rowe, M. (1974) Relation of wait-time and rewards to the development of language, logic and fate control: Part II, *Journal of Research in Science Teaching*, 11(4) 291-308

[63] Hattie, J. and Timperley, H. (2007) The power of feedback. *Review of Educational Research*, 77, 81-112.

The form of the feedback

It has been traditional for a good deal of feedback on students' work to be in the form of a grade, mark or other sign of a judgement of its adequacy. A frequently cited research study by Butler[64] into feedback compared giving marks with two other forms, one giving comments on the work and how to improve it, and the other giving such comments and marks. This was a well-designed and complex study which involved different kinds of task and students of different levels of achievement. One of the outcomes was that feedback in terms of comments-only led to higher achievement for all students and all tasks. An interesting result was that providing both comments and marks was no more effective than marks alone. It appears that students seize upon marks and ignore any comments that accompany them. They look to the marks for a judgement rather than help in further learning. When marks are absent they engage with what the teacher wants to bring to their attention. The comments then have a chance of improving learning as intended by the teacher. In order to do this, of course, the comments should be positive, non-judgemental and where possible identify next steps.

In their work on implementing formative assessment, Black et al (2003) introduced teachers to Butler's research and to the results of observation in their own classrooms, which showed that:

- Students rarely read comments, preferring to compare marks with peers as their first reaction on getting work back
- Teachers rarely give students time in class to read comments that are written on work and probably few, if any, students return to consider them at home
- Often the comments are brief and/or not specific, for example 'Details?'
- The same written comments frequently occur in a student's book, implying that students do not take note of or act on the comments.[65]

When students were asked about the way in which their books were marked, they wanted teachers not to use red pen, to write legibly and to make statements that could be understood.

The content of the feedback

The main point to emerge both from research studies and from experience of effective practice is a distinction between feedback that gives information and feedback that is judgemental. This applies to oral as well as written feedback. Feedback that gives information:

- focuses on the task, not the person
- encourages students to think about the work not about how 'good' they are
- indicates what to do next and gives ideas about how to do it.

Feedback that is judgemental:

- is expressed in terms of how well the *student* has done (which includes praise as well as criticism) rather than how well the *work* has been done
- gives a judgement that encourages students to label themselves
- provides a grade or mark that students use to compare themselves with each other.

A further aspect of effective feedback is that it focuses students' attention on what is relevant to achieving particular lesson objectives. Thus if the main focus of an inquiry is to develop science inquiry skills, then feedback should be about that aspect rather than, say, the particular results obtained.

[64] Butler, R. (1988) Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest and performance, *British Journal of Educational Psychology* 58, 1-14

[65] Black et al, (2003) op cit p 43

At the start of their work with teachers, Black et al noted that the comments made by teachers on students' work were either a general evaluation ('Good', 'Well done') or focused on the presentation or completeness of the work. One of the teachers, recognising that such remarks did not help the students' learning, wrote in his notes:

> A bland, non-helpful comment such as "Good work, Jaspaul, this is much neater and seems to show that you have tried hard" will not show a significant change in attainment because it says nothing about the individual's learning. There is no target and the student, although aware that the teacher is happy with them, could not be blamed for thinking that neatness is everything...Students are not good at knowing how much they are learning, often because we as teachers do not tell them in an appropriate way.[66]

Comments that help learning are such as:

> "You have correctly identified which chemicals are elements and which are compounds, now try to give a general explanation of the difference between elements and compounds."
>
> "Look back at your notes on cell structure and then see if you agree with what you have written here."
>
> "What did you find that made you decide that the mass of the objects makes no difference to how quickly they fell?"

Conclusions about feedback to students

Recommendations that come from research are that if feedback is to help learning, then:

- It should be in the form of comments with no marks, grades or scores.
- Whether oral or written, comments on students' work should identify what has been done well, what could be improved and how to set about the improvement.
- Comments should help students to become aware of what they have learnt.
- Teachers should check that students understand their comments.
- Time should be planned for students to read and, if appropriate, respond to comments.

Feedback into teaching

Formative assessment is as much about feedback to teachers as it is about feedback to students. The two are closely related, for how students respond to the questions and feedback from their teacher and from other students is a source of evidence for the teacher to use in making decisions about the next steps for the students. Since all learners are individuals, there is no single path to follow to ensure learning. Teachers have to judge the value of an intervention from the impact of their questioning and other actions. In order to collect relevant data to inform their interventions, teachers need to be very clear about the goals they want their students to achieve. Then they can distinguish the significant data from the vast range that is potentially there to be used. With this focus they will be able to collect data when students are involved in investigations, by observing, questioning, listening to how students are using words and studying notebooks. An important source of feedback to the teacher comes from students' self-assessment and peer assessment, discussed below, for if the students are not clear about what they should be trying to achieve through their work, they may be using inappropriate criteria for judging success.

[66] Black et al, (2003) op cit p 45

This feedback informs teachers' decisions about whether or how to intervene in the course of students' activities. The process is cyclical and each decision changes the situation. Not all interventions will have the desired positive impact; what happens in classrooms does not always – perhaps rarely – goes as planned. The feedback teachers receive from students' reactions enables them to try something different, if necessary, in order to help students to make progress. It may be necessary for a teacher to change plans when students are struggling rather than risk a sense of failure. In this way the feedback enables teachers to regulate teaching to maximise learning.

Student self- and peer-assessment

A common goal of formative assessment and IBSE is that students become increasingly able to take part in decisions about the quality of their work and develop their understanding of what is involved in learning. Learners are, in any case, responsible for learning, but whether they *take responsibility* for it depends on their participation in decisions. This participation is represented by the two-headed arrows in Figure 1 page 18.

Students, like all learners, direct their effort more effectively if they know what they are trying to achieve, rather than just knowing what they have to do. Teachers are good at telling students what to do ('wrap the ice cubes in different materials and see which takes the longest to melt') but not so good at providing a purpose and a goal ('to see if some materials are better than others for keeping ice from melting and try to explain what you find'). A prerequisite for being able to judge their work is that students understand what they are trying to do, not in terms of what is to be found, but in terms of the question to be addressed or problem to be solved.

Of course, if the question or problem is one raised by the students, the need to communicate it does not arise. But this will be only one of a range of pedagogical practices used by teachers, including teacher-directed activities. There will always be situations where teachers introduce the question and, although using skills to help students 'make it their own', teachers will need to make sure that students understand the purpose and goal of the activity. Stating goals at the start of a lesson is not the only, or necessarily the best, way of conveying them. The understanding of the goals, of why they are working in a particular way can be reinforced by dialogue and questions during the activity and in discussion of what was done and found.

In order to assess their work, students not only need to know the purpose of what they are doing but they need to have some notion of the standard they should be aiming for, that is, what is 'good work' in a particular context. Some of this is conveyed implicitly through the feedback that teachers give to students. Teachers can also discuss more explicitly what makes one piece of work or investigation better than another, using examples collected for the purpose and made anonymous. Alternatively, the examples from the collections published or created in the school to help teachers assess work could be shared with the students (see later). Box 15 describes some examples of how teachers of primary students have approached the discussion of what is good work.

With older students more direct and sophisticated approaches can be used to encourage self-assessment. Box 16 describes how a biology teacher helped his 16 year old students to an understanding of how to evaluate their work. It arose in the course of helping the students to clarify their ideas about plant nutrition.

Box 15: Communicating to primary school students criteria for evaluating their work [67]

Using examples

A teacher of 10 year olds, spent some time at the beginning of the school year discussing with her class what made a 'good' report of a science investigation. She gave students two anonymous examples of students' writing about an investigation produced by other students in earlier years. One was a clear account, well set out so that the reader could understand what had been done, although the writing was uneven and there were some words not spelled correctly. There were diagrams to help the account, with labels. The results were in a table, and the writer had said what he or she thought they meant, admitting that the results didn't completely answer the initial question. There was a comment about how things could have been improved. The other account was tidy, attractive to look at (the diagrams were coloured in but not labelled) but in content contained none of the features shown in the other piece.

The teacher asked the students to work in groups to compare the pieces of work and list the good and poor features of each one. Then they were asked to say what they thought were the most important things that made a 'good' report. The teacher put all the ideas together and added some points of her own, to which the students agreed. She later made copies for all the students to keep in their science folders. But she also went on to explore with them how to carry out an investigation in order to be able to write a good report. These points too were brought together in the children's words and printed out for them.

Brainstorming

A variation on the above is to brainstorm their ideas about, for example, how to conduct a particular investigation so that the children are sure of the result. The list of what to think about can be turned into questions (Did we keep everything the same except for …? Did we change …? Did we look for…? Did we check the results? etc.). Before finishing their investigation they check through their list, which becomes a self-assessment tool for that piece of work.

Discussing 'best work'

This approach can be used with students from about the age of 8. It begins with the students selecting their `best' work to put into a folder. Time is set aside for the teacher to talk to each student about why certain pieces of work have been selected. During this discussion the way in which the students are judging the quality of their work will become clear. These are accepted without comment, whether or not they reflect the teacher's view of good work. To clarify the criteria the students use, the teacher might ask. `Tell me what you particularly liked about this piece of work?' Gradually it will be possible to suggest criteria without dictating what the students should be selecting. This can be done through comments on the work. `That was a very good way of showing your results, I could see at a glance which was best.' `I'm glad you think that was your best investigation because although you didn't get the result you expected, you did it very carefully and made sure that the result was fair.'

[67] Adapted from Harlen, W. (2006) *Teaching, Learning and Assessing Science 5-12*. London: Sage. p171

Box 16: Secondary students using self-assessment to improve their work [68]

The students were given the following question to discuss: 'If a villain got hold of a chemical that could destroy chlorophyll, what effect would this have on plants if released?' Each group of four were asked to write down between three and five criteria that they felt needed to be met by a good written answer to this question. These criteria were discussed by the whole class and a final list was drawn up:

- Say what chlorophyll does
- Explain photosynthesis as a process
- Put in the equation for photosynthesis
- Describe the effect on plants of no chlorophyll
- Add any secondary effects from stopping photosynthesis.

The students then wrote up their answers for homework, which the teacher marked, writing comments only on their work. Then in pairs, students read the comments and each other's work to check that they understood what the teacher was asking them to do to improve. They were then given time in the lesson to redraft and improve their answers.

Peer-assessment

Peer assessment by students has been given a central role in formative assessment, following its strong advocacy by Sadler.[69] Black provides several arguments in favour of encouraging students to judge each others' work. Some of these are based on helping students to understand better the goals and criteria of quality by looking at another's work rather than their own. Some reasons are based on practicality – that students will be motivated to improve their work if they know that it will be read by another student.[70] However, as other researchers have pointed out, these arguments do not take account of the power relationships associated with one person judging another's work.[71] Case studies by Crossouard of peer assessment by students aged 11 and 12, provides disturbing evidence of how gender, social class and students attainment hierarchies are 'implicated in processes that are typically bathed in the supposed 'neutrality' of assessment judgements.'[72] The benefits of peer assessment were observed to be unequally spread, the practice supporting and extending some students whilst working 'oppressively' for others. Thus teachers may need help in recognising the issues of equity that are raised and in addressing the influence of social class, gender and general ability when practising peer assessment.

[68] Based on Black et al (2003) op cit p 63

[69] Sadler (1989) op cit

[70] Black et al (2003) op cit p 50

[71] Crossouard (2012) Absent presences: the recognition of social class and gender dimensions within peer assessment interactions, *British Educational Research Journal*, 38 (5) 731-748
Pryor, J. and Lubisi, C. (2001) Reconceptualising educational assessment in South Africa –testing times for teachers, *International Journal for Educational Development*, 22 (6), 673-686

[72] Crossouard (2012) op cit p 736

Chapter 6
Implementing summative assessment of IBSE

After considering, in Chapter 5, some approaches to putting formative assessment into practice to support students' learning through inquiry, in this chapter we look at some methods of implementing summative assessment. The aim is to provide information about what students understand and can do at a certain time with an emphasis on scientific knowledge and understanding and science inquiry skills. By contrast with the requirements of formative assessment, reliability is important in summative assessment since the results may be used to compare or select students and may also be used as indicators of teacher and school effectiveness. However, validity is also important and, given the interaction between validity and reliability noted in Chapter 1 (Box 2), a key factor in choosing assessment methods is to ensure that validity is not compromised in striving for apparent accuracy. We focus here on summative assessment of scientific understanding and science inquiry skills since these constitute the greatest challenge and their neglect in assessment the greatest threat to IBSE.

This focus does not mean that other outcomes, such as factual knowledge, scientific vocabulary and use of conventions are not to be assessed, but there are already many familiar ways of doing this – for instance, through short classroom tests and quizzes devised by the teacher at appropriate times. More challenging is the assessment of progress towards the understanding and competences that are the goals developed through IBSE. Unless these are included in the assessment there is a real danger that they are neglected in teaching.

Assessment of understanding and inquiry skills

Understanding and 'big' ideas

As a start we need to consider what understanding is and how it is related to specific and general knowledge. Understanding is not easily pinned down, for as White points out:

> (understanding) is a continuous function of a person's knowledge, is not a dichotomy and is not linear in extent. To say whether someone understands is a subjective judgement which varies with the judge and with the status of the person who is being judged. Knowledge varies in its relevance to understanding, but this relevance is also a subjective judgement.[73]

Understanding exists at various levels and in different ways of explaining phenomena. The understanding of a primary school student and a higher education science student will be different but both may meet the expectations appropriate to their stage of learning. For example, a young student might explain the phenomenon of dissolving in terms of what happens when some solids are added to some liquids. Older students will be expected to explain it in terms of molecular behaviour that can apply to solutions of gases and liquids in liquids. What is probably the same for both is that the ideas they have make sense to them and fit the experiences they are trying to understand at a particular time. As the younger students' experiences expand and their ideas no longer provide a satisfactory explanation, so the phenomenon will need to be understood in a different way. Progressive understanding that accompanies expanded experience can also be seen as the development of successively 'bigger' ideas, linking more phenomena and being more powerful in explaining things.

[73] White, R.T. (1988) *Learning Science*. Oxford: Blackwell. p 52

'Big' ideas are ones that can be applied in different contexts; they enable learners to make sense of a wide range of phenomena by identifying the essential links or 'meaningful patterns' between different events, objects or phenomena without being diverted by differences in superficial features. They can be ideas of science (such as about forces and movement) or about procedures of science (about manipulation of variables). Merely memorising facts or a set of procedures does not support this ability to apply learning to contexts beyond the ones in which it was learned. Knowledge that is understood is thus useful knowledge that can be used in problem-solving and decision-making.

Thus progress in understanding shows not in the amount of knowledge but whether it is organised in the mind, forming 'big' ideas, so that it can be easily accessed and applied to new experience. This demonstration of understanding is a key aspect of performance that distinguishes 'experts' from 'novices'.

> Experts have acquired extensive stores of knowledge and skill in a particular domain. But perhaps most significant, their minds have organised this knowledge in ways that make it more retrievable and useful.
>
> …These methods of encoding and organising help experts interpret new information and notice features and meaningful patterns of information that may be overlooked by less competent learners. These schemas also enable experts, when confronted with a problem, to retrieve the relevant aspects of their knowledge.[74]

Here, then, are some clues to assessment in science that afford inferences to be drawn about students' understanding:

- The assessment task has to involve application of knowledge, not only recall.
- This means that the task has to be novel but at the same time must not be so far from their experience that it has no meaning for the students.
- In recognition that science is about the real world, not an imagined one, the task must be authentic, about real things and real data.

There are other requirements, too, if the assessment task is to be presented within a reasonable time and not depend too much on competences (such as reading and writing ability) other than the understanding being assessed.

Science inquiry skills

In all assessment tasks that are set in contexts and dealing with subjects that are meaningful to students, performance will be affected by students' familiarity with the context and subject matter. All competences are used in a context and in relation to some content. Although the context also has an influence on application of science concepts, it presents particular problems when the focus of the question is to elicit data about use of inquiry skills.

Questions and tasks have to be asked about *something*; observations are made about particular objects and events; investigations are planned to answer questions about particular phenomena; there has to be some subject matter involved when skills are used. What this subject matter is makes a difference to whether skills are used. A student may be able to plan an appropriate investigation about a situation where (s)he has some knowledge of what are likely to be the variables to control, but fail to do this if the subject matter is unfamiliar. This has important consequences for assessment. The subject of a particular task or test item is just one of a potentially large number of alternative subjects. In theory a student's result would be different if a different subject had been chosen, thus there is a variation, or error, in the results associated with the choice. This is described as the

[74] Pellegrino et al op cit p 72/73

'sampling error'. It is an unavoidable error since no two tasks with different subject matter or contexts can be exactly equivalent.

Some points about assessment of science inquiry skills follow from this:

- First, is the obvious one that students need to be involved in using the inquiry skills in order to assess what they can do.
- Second, since the context and subject of the situation in which the skills are to be used affects the ability to use the skills, the task should be set in a familiar context if possible or several contexts should be used to reduce the sampling error.
- Third, as in the case of assessing understanding, the tasks should be authentic and engaging to the students.

The task they are undertaking may involve

- carrying out in practice a complete inquiry to address a given question or problem providing opportunity for evidence to be collected about the use of a range of the skills
- producing a plan on paper for a complete inquiry to address a given question or problem
- considering a particular part of a given investigation, such as the variables that need to be manipulated or controlled, the evidence that needs to be collected, or the interpretation of some given data.

Methods of summative assessment of IBSE goals

This brief discussion of what is needed to assess understanding and science inquiry skills shows that in both cases students need to be working on tasks where some aspects of inquiry are involved. Both also require some novelty in the task that students are undertaking so that they are using their knowledge or skill and not simply recall of information, reasons or procedures that have been committed to memory. For example, the task of finding out how the thickness of a conducting wire affects its resistance is not one that is going to assess ability to plan and conduct an investigation if it is something that students have already done, often more than once. Neither will a question about why the age of a tree can be estimated from its growth rings probe understanding of ideas about growth of trees if the reason has been memorised without being linked to these ideas.

For understanding, the task should require an explanation of an event or interpretation of data or a prediction involving application of some concepts. For skills, the task has to be accomplished by using one or more of the inquiry skills, such as predicting, planning, carrying out an investigation or interpreting given data. However, as already noted on page 12, it is not possible to assess skills without involving some knowledge of the subject matter of its use. At the same time, tasks used for assessing understanding will require some use of skills (explaining, interpreting, predicting). Thus there will always be some aspects of understanding and skill required in *all* tasks. What determines whether a task is essentially assessing understanding or skill will be the level of demand on one or the other, and the credit given to different kinds of responses in scoring.

Turning to methods of assessment of these goals, we first consider what can be done through testing, since this is the most common approach to summative assessment. Later we look at alternatives to tests.

The potential and limitations of tests for assessing IBSE goals

Assessment for summative purposes needs to be as reliable (free from error and bias) as possible. There is an attraction in using special tasks or tests because they can be controlled and presented to all students in the same way, thus appearing to give the same opportunities for students to show what they can do. Testing, as noted in Chapter 1, is a method of assessment in which procedures, such as the task to be undertaken and often the conditions and timing for responding are specified. Tests are usually marked (scored) using a prescribed scheme (rubric), either by the students' teacher or external markers, who are often teachers from other schools, or by machines. The reason for the uniform procedures is to allow comparability between the results of students, who may take the tests in different places. There are different forms of test according to the nature of the task and the form in which a response is given. So there are

- performance tests (sometimes called 'practical')
- embedded tests (set in the context of regular work)
- multiple choice tests (where alternative answers are provided)
- open-ended, or open-response tests (where student write short or long answers in their own words)
- open-book tests (where students have access to a controlled number of sources)

… and many more. Most involve some writing, except in some performance tests and tests of reading for young children. The more formal tests, leading to a certificate or qualification, are often described as examinations.

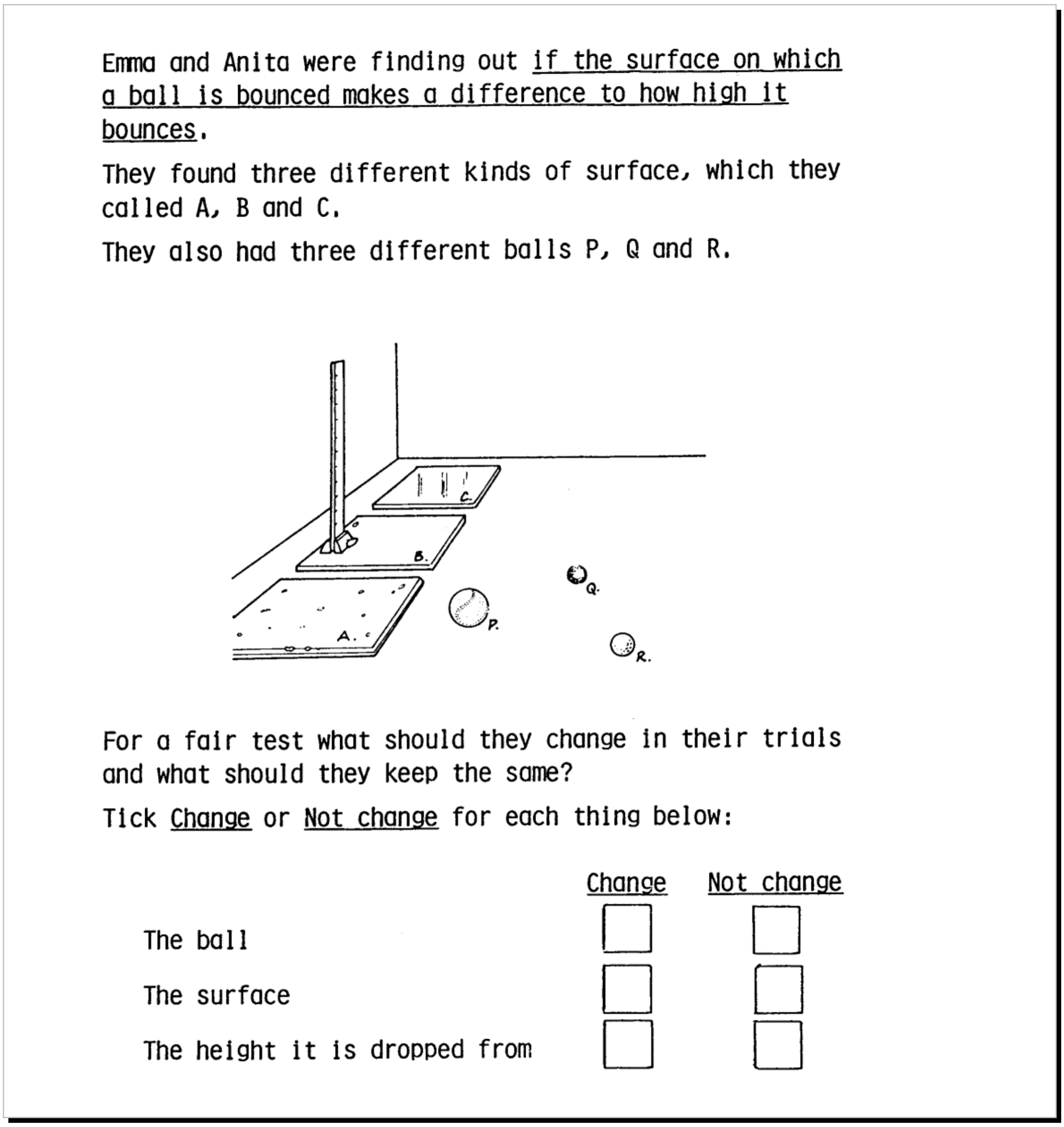
The potential of tests for assessing scientific understanding and science inquiry skills is best explored through a few examples of written and performance tests.

Written items

The features of good quality oral questions noted in chapter 5 apply equally to written questions. That is, in order to gain access to students' thinking, it is useful to express questions in terms of what students think. 'Open' questions, asking for students to respond in their own words are also more likely to elicit information about what they know and understand than questions where students choose between given answers. These are open to guessing and clues given by the words used in the options. However, open response questions also have disadvantages, particularly for younger students who have greater difficulty in being specific in writing than in talking. Interpreting students' written answers can also be a problem, although there are now computer programs that can do this with considerable reliability.[75] Further, it is necessary for a question to make clear to the students what kind and extent of answer is required. For instance to ask: 'What are the main differences between X and Y?' may lead to a long list of differences, some relevant, some not and some inaccurate. It is better to make clear what kind of answer is required, as in: 'Write down what you think are the three most important differences between X and Y that enable them to survive in different habitats'. Then to provide spaces for these answers that indicate limits to the length of the answers. The advantages and disadvantages of different question formats means that a mixture is usually included, the balance of each type varying according to the aims of the test, the particular focus of a question and the age of the students. Some further points emerge from considering some examples.

[75] Streeter, L. et al (2011) *Pearson's Automated Scoring of Writing, Speaking, and Mathematics*: Pearson

Example 1: ***Bouncing balls*** *(From APU Report of Performance at age 11)*[76]



Emma and Anita were finding out if the surface on which a ball is bounced makes a difference to how high it bounces.

They found three different kinds of surface, which they called A, B and C.

They also had three different balls P, Q and R.

For a fair test what should they change in their trials and what should they keep the same?

Tick Change or Not change for each thing below:

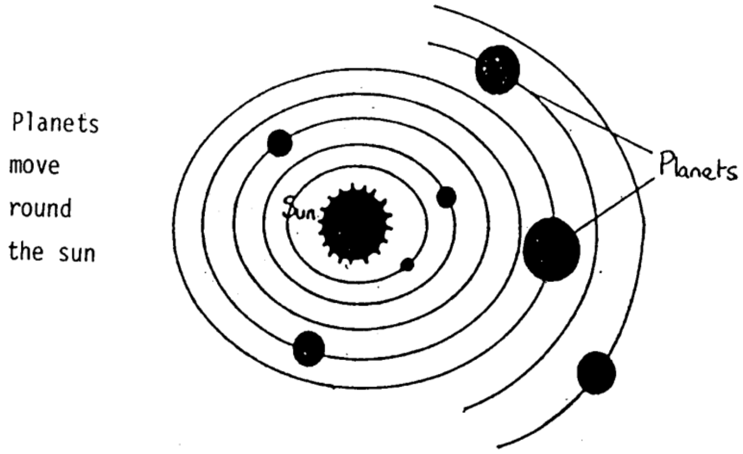
	Change	Not change
The ball	□	□
The surface	□	□
The height it is dropped from	□	□

Comment

Example 1 (Bouncing balls) requires students to identify with the situation described and the question being addressed. The subject matter is likely to be familiar to all students, thus the level of knowledge required is low and the burden of the task is about conducting a fair test. The format for answering, and the requirement of the scoring scheme for the answer in each box to be correct, makes the chance of succeeding by guessing very low. This is one way of making a multiple choice question into one that reduces the chance of success by guessing. It also means that students have to read and understand the instructions for recording their answer, otherwise there is a risk of failure for reasons other than not having the skill needed to the answer the question.

[76] DES, DENI and WO (1985) *APU Science in Schools Age 11 Report no 4*. London: HMSO

Example 2: ***Planets*** *(From APU Age 11 Report No 1)*[77]



Look at the following table

Planet	Distance from the Sun	Time for one trip round the Sun
Mercury	**58 million kilometres**	**88 days**
Venus	**108 million kilometres**	**225 days**
Earth	**150 million kilometres**	**1 year**
Jupiter	**780 million kilometres**	**12 years**
Uranus	**2870 million kilometres**	**84 years**
Neptune	**4500 million kilometres**	**165 years**

a) There is another planet not in this table. It is about 1430 million kilometres from the Sun. About how long do you think it will take this planet to make one trip round the sun?

- □ **10 years**
- □ **100 years**
- □ **100 days**
- □ **30 years**
- □ **300 years**

b) Why do you think it will take this time?

Because………
……

Comment

In Example 2 (Planets) all the information required to answer the question is given and pupils do not need to know anything about planets. The intended focus of the task is to recognise the pattern between the two sets of figures given. However, presented 'cold' to a student (that is, not as part of a study relating to the solar system) it may appear to require more knowledge than is the case. Some students may not respond if their first impression is that they need to know about planets. The scoring can be set so that some credit is given for part a) even though there is a one in five chance of guessing correctly, but more for part b) according to whether or not the pattern is stated explicitly. The open-ended nature of part b) is more demanding than selection from given alternative reasons. It provides more information about how students interpret data but is less easily marked.

[77] DES, DENI and WO (1981) *Science in Schools Age 11 Report No 1.* London: HMSO

Example 3: ***Climate Change*** *(From the PISA assessment of Science 2000)*[78]

Read the following information and answer the questions which follow.

WHAT HUMAN ACTIVITIES CONTRIBUTE TO CLIMATE CHANGE?

The burning of coal, oil and natural gas, as well as deforestation and various agricultural and industrial practices, are altering the composition of the atmosphere and contributing to climate change. These human activities have led to increased concentrations of particles and greenhouse gases in the atmosphere.

The relative importance of the main contributors to temperature change is shown in Figure 1.

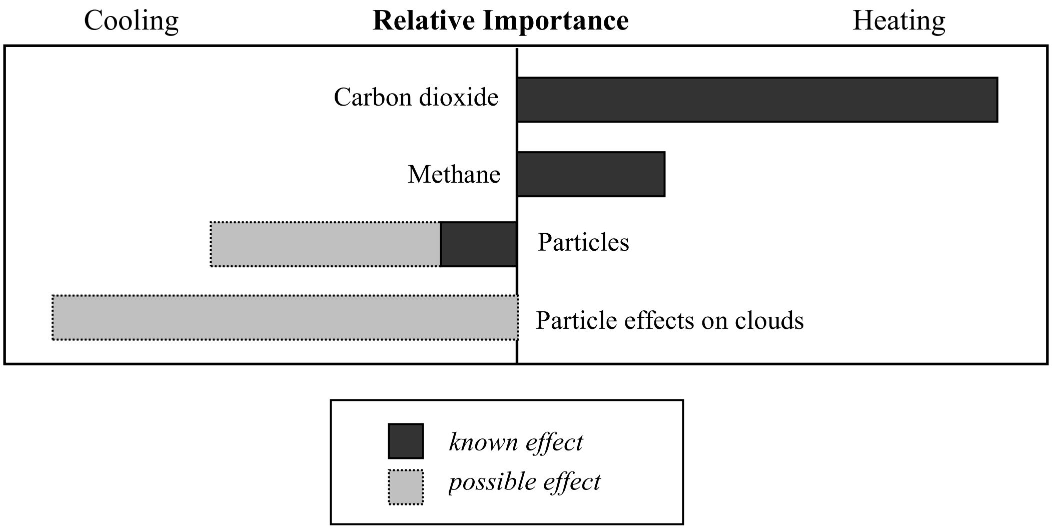
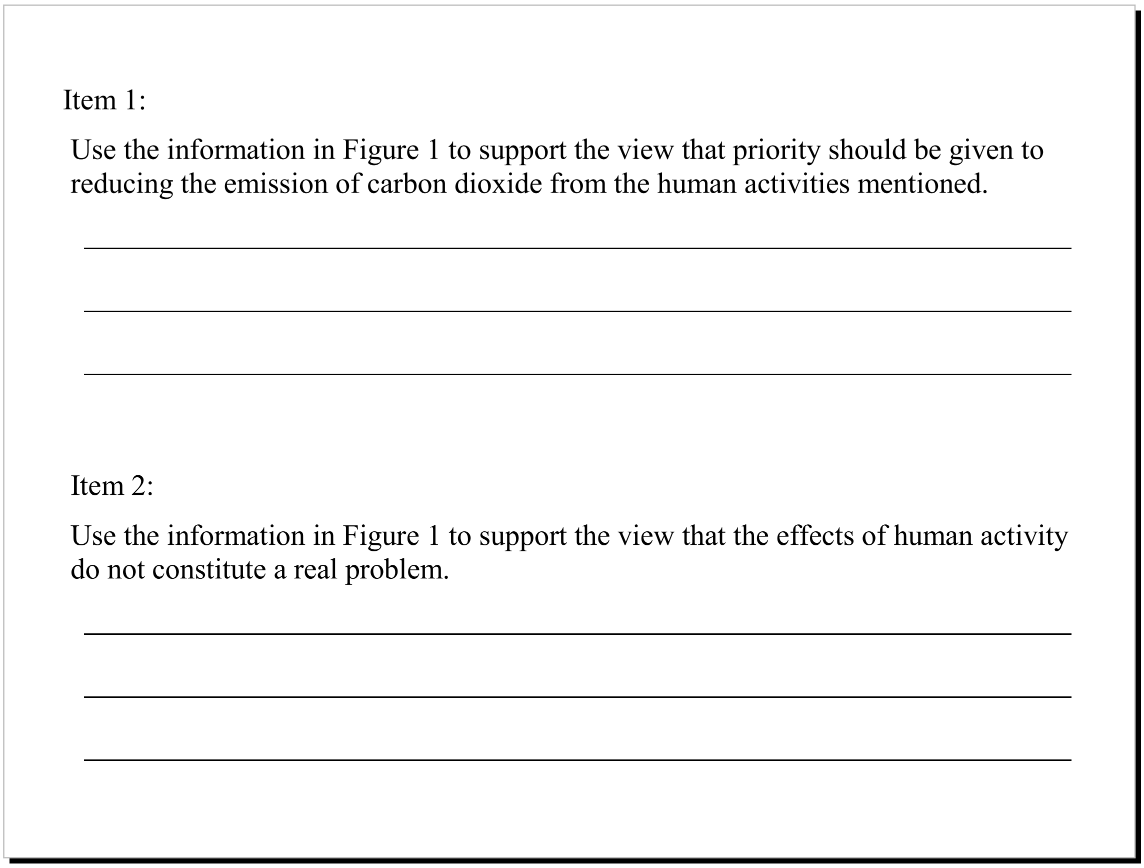


Figure 1: Relative importance of the main contributors to change in temperature of the atmosphere. *Source: adapted from http://www.gcrio.org/ipcc/qa/04.html*

Bars extending to the right of the centre line indicate a heating effect. Bars extending to the left of the centre line indicate a cooling effect. The relative effect of 'Particles' and 'Particle effects on clouds' are quite uncertain: in each case the possible effect is somewhere in the range shown by the light grey bar.

Figure 1 shows that increased concentrations of carbon dioxide and methane have a heating effect. Increased concentrations of particles have a cooling effect in two ways, labelled 'Particles' and 'Particle effects on clouds'.

[78] OECD (2000) op cit p 88



Item 1:

Use the information in Figure 1 to support the view that priority should be given to reducing the emission of carbon dioxide from the human activities mentioned.

____

____

____

Item 2:

Use the information in Figure 1 to support the view that the effects of human activity do not constitute a real problem.

____

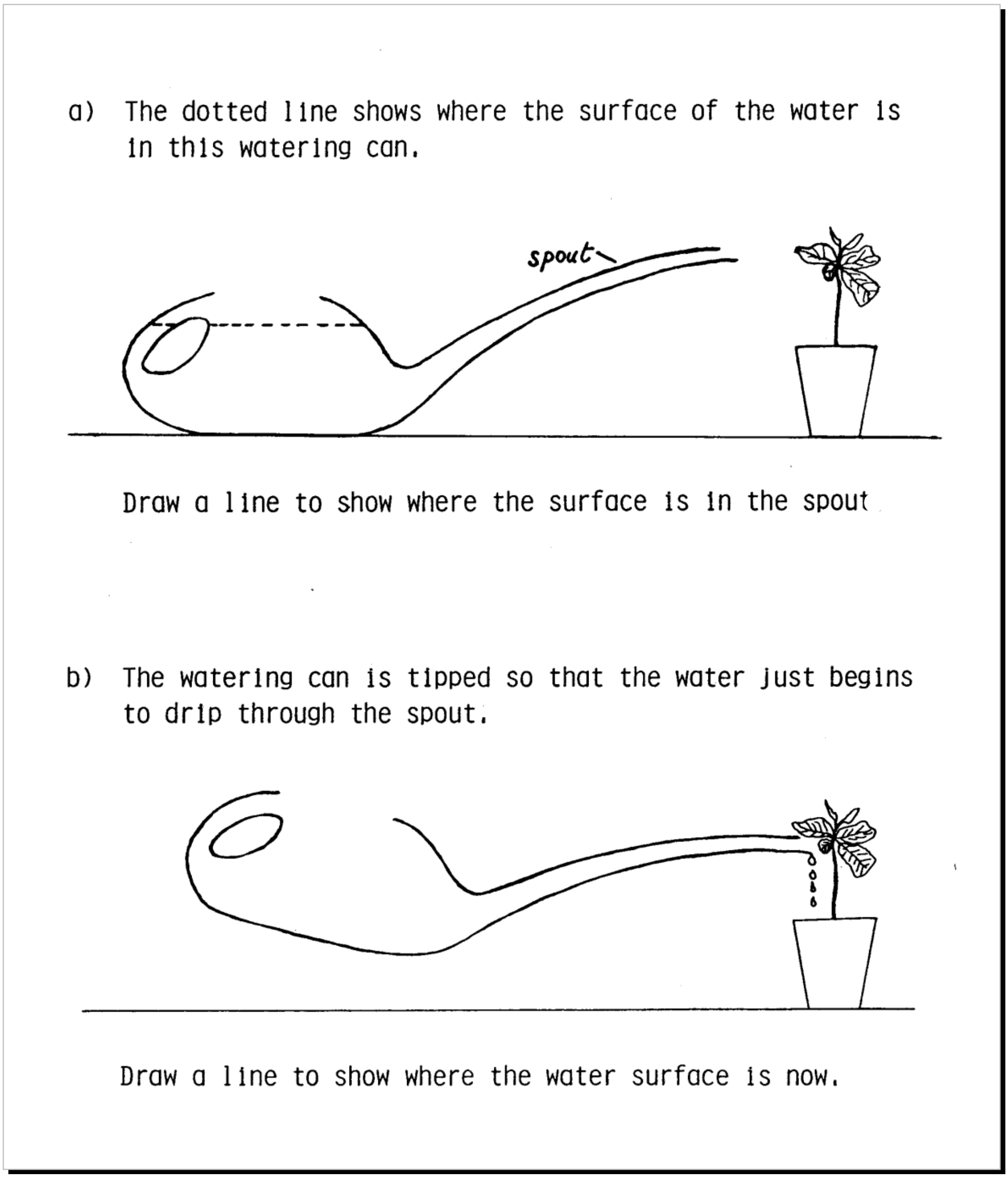
____

____

Comment

Example 3 (Climate change) is another item where information is given and students are asked to use it, in this case to support alternative conclusions about action that the data suggest could be taken. The information is authentic and presents the sort of problem that the scientifically literate should be able to engage with. The two parts to the task illustrate the uncertainty of interpreting scientific information in certain cases. In theory all the information is provided and the students are told how to interpret the graph. They do not need to know how carbon dioxide, methane, particles and their effects on clouds cause heating and cooling. However, without any knowledge of these things the question is likely to be meaningless and they are unlikely to engage with the problem posed.

*Example 4: **Watering can** (From APU Age 11 Report No 1)*[79]



a) The dotted line shows where the surface of the water is in this watering can.

Draw a line to show where the surface is in the spout

b) The watering can is tipped so that the water just begins to drip through the spout.

Draw a line to show where the water surface is now.

Comment

Example 4 (Watering can) requires application of knowledge about water flowing until reaching a common level. The context is assumed to be familiar and the response is by drawing, thus being less dependent on writing and vocabulary than other items. However, although a scoring scheme can allow for inaccuracy in drawing, the correct completion of the task still requires students to read how they are intended to respond.

[79] DES, DENI and WO (1981) op cit p 98

General comments on written test items

There are several points that emerge from these examples. The most obvious is the inevitable demand for reading and understanding the question and, depending on the answer format, for writing ability. Then the attempt to place the task in a context that can seem real to the student means that some sort of 'story line' is presented as a context for the task. Students have to read and engage with the context in order to respond to the question. We need to ask, then, whether the nature of this context makes a difference to how the student responds. For instance, would the identification of variables in Example 1 be more difficult for some students if the context were about comparing the effect of changing the ingredients in making a cake, or the speed of toy cars down a slope? Would the ability to identify a pattern in data be affected by whether the data are presented pictorially rather than as numbers? Does the answering format affect students' performance? The answer to all these questions from research is that these features do matter and the context of questions is particularly important to students' ability and willingness to engage and show what they can do. Other circumstances which raise matters of fairness, particularly for students in countries with diverse populations, are noted in Box 17.

Box 17: Unequal opportunities in tests

Research shows that familiarity with the context in which test items are set is a factor affecting students' performance. When familiarity with content is unequally spread among the population, this raises questions about the equity of the test. Opportunities for the relevant experience needed to understand a test item may be denied to some students, for example to girls or to those from economically poor backgrounds. Some students may struggle with the language of the test[79] when this may be their second or third language. These students, and those with special physical needs or learning difficulties, may not be given the opportunity to show what they can do and as a consequence may be denied access to further education or training.

In order for the effect of a particular context or format to be minimised, as large a range of contexts and formats as possible should be included in a test. Since there is a limit to the length of test a student can be expected to undertake, the consequence is that there is a tendency to favour including more short items which can be spread over several contexts, and more closed items which can be answered by selection rather than giving an answer in their own words, which have the additional advantage of being able to be marked by machine.

It is important to note that the restriction on length and thus on range of contexts is most severe when test individual students are all given the same test items. It is not as great a problem in the case of a population survey where a large number of items can be used, divided into a number of sub-tests each given to a random sample of the population. This is the design used in national surveys such as the NAEP and the APU and international surveys of TIMSS and PISA. The purpose of these surveys is to identify performance at national or regional levels, which is derived from the combination of results from the sub-tests. The results for individual students, or even for a whole school, in the sample have little value in terms of the overall goals being assessed, but when combined across the sample, the result provide a much better picture of national performance than would be obtained if every student took the same test items. When different items are given to different student samples, there is no need to cram into a short test a number of short items and more time can be given for students to engage with a particular context. It is therefore not a coincidence that the examples cited here were created and used in APU and PISA programmes.

[80] Noble et al (2012)) "I never thought of it as freezing": How Students Answer Questions on large-scale science tests and what they know about science*, Journal of Research in Science Teaching*, 49 (6) 778–803.

Box 18: Sources of science assessment test items and tasks

SEAR (Science Education Assessment Resources)[81]

Comprises a bank of assessment resources which cover the full age range of schooling from year 1 to year 12. Tasks are labelled, and can be searched and accessed according to scientific literacy level, assessment purpose, task type and learning outcome focus. Six levels of scientific literacy are defined by a 'Scientific Literacy Progress Map' which is linked to the PISA framework. The bank includes tasks for diagnostic, formative and summative purposes, and are in multiple choice, open ended and practical formats. Extensive mark schemes accompany each item.

PISA (Programme for International Student Assessment) released science items[82]

This is a collection of the items that have been used in the surveys of pupils at age 15 and published in various reports. Each item includes the stimulus material and the scoring scheme.

APU (Assessment of Performance Unit) science items

Examples of written and practical test items and tasks used in survey of students aged 11, 13, and 15 in England and Wales in the 1980s are available in the reports written for teachers.[83] The marking scheme and a description of students' responses is given for each item.

NAEP (National Assessment of Educational Progress – the Nation's Report Card) Science Sample Items 2011[84]

Provides all the questions released from the 2011 science assessment for grades 4, 8, and 12. Items include multiple-choice, short response and extended response types. For each question, scoring guides, students response and performance data are provided.

Box 18 lists a number of sources of published written and performance items which have been used in science tests and surveys. They include, but are not restricted to, items which assess application of knowledge and understanding and the use of inquiry skills relevant to assessing the outcomes of IBSE.

Performance items

Two of the deficiencies of written tests – the dependency on reading and writing skills and the need to engage with a problem in context which is only presented on paper – can be avoided to some extent if the student is able to undertake an inquiry with real objects and equipment. There are, of course, other restrictions. The question still has to be presented to the student, who has to engage with it as if it were his or her own, and the situation is far from that of a normal classroom, since the students may be working alone (sometimes in pairs) with an administrator present to observe their actions. However, it does give an opportunity for student to explore, try things, and start again if necessary. Example 5 shows an investigation about Paper Towels that was used in surveys of students aged 11, 13 and 15 in the APU surveys in England in the 1980s.

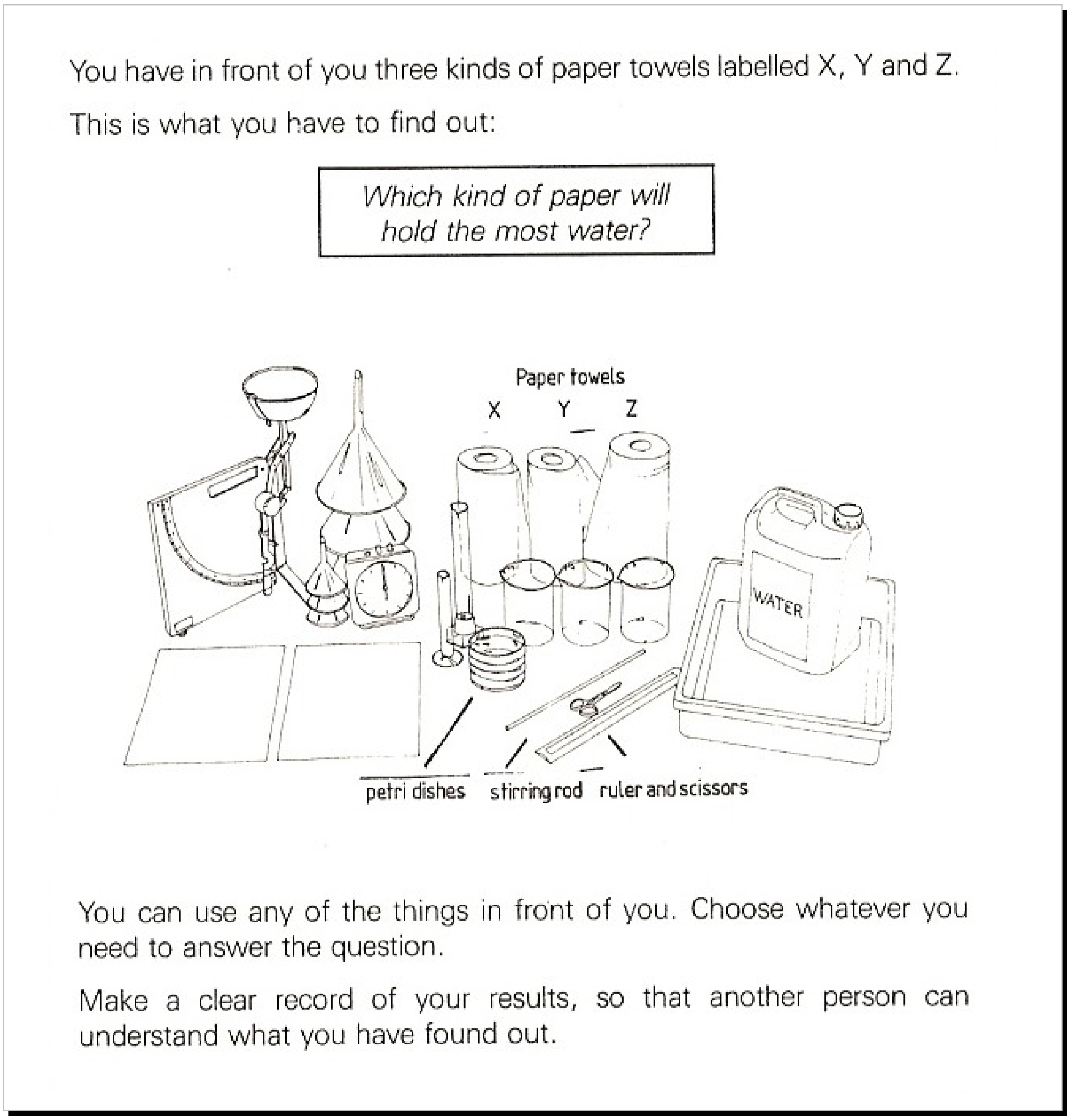
[81] Australian Government Department Education, Employment and Workplace Relations. *Science Education Assessment Resources (SEAR)* http://cms.curriculum.edu.au/SEAR

[82] OECD (2006) PISA released items: Science. Paris: OECD http://www.oecd.org/pisa/38709385.pdf

[83] http://www.nationalstemcentre.org.uk/elibrary/collection/727/assessment-of-performance-unit-science-reports-for-teachers

[84] US Department of Education http://nationsreportcard.gov/science_2011/sample_quest.asp

Example 5: ***Paper towels*** *(From APU Science report for teachers no 6, 1985)*[85]



You have in front of you three kinds of paper towels labelled X, Y and Z.

This is what you have to find out:

Which kind of paper will hold the most water?

Paper towels
X
Y
Z
WATER
petri dishes
stirring rod
ruler and scissors

You can use any of the things in front of you. Choose whatever you need to answer the question.

Make a clear record of your results, so that another person can understand what you have found out.

Comment

This is one of the tasks administered to individual students working alone, by a trained tester. The equipment is all provided by the test programme and is set out in a uniform manner before a student enters the room. The student is introduced to the equipment and given the opportunity to explore it. The tester hands the student a paper on which the question is written, with spaces for notes and a record of results and checks that the student understands the question. During the investigation the tester completes a detailed checklist, intervening only if safety is threatened. Finally when the student has finished the practical task, the tester looks at what the student has written to ensure it can be understood and probes the evidence for the answer. The test ends by the student being asked what change, if any, they would make if they could begin again.

[85] Welford, G., Harlen, W. and Schofield, B. (1985) *Practical Testing at ages 11, 13, and 15*. London: DES, WO and DENI

Reducing the error due to some factors increases the influence of others, in particular the effect of the context. In a test session of reasonable length, any one student can only undertake a small number (two or three at most) of such investigations. Even in a national survey there is a limit to the total number that can be used and the effect of the context defeats any attempt to provide a score on 'performance of investigations' derived from performance in several investigations. The surveys in which they are used provide the evidence that students who perform well in one investigation will not necessarily do so in another testing the same skills but in a different context.

This was confirmed in a carefully designed study in the USA. Pine et al (2006) assessed fifth grade students using several 'hands-on' performance tasks, including one based on 'Paper Towels' and one called 'Spring', about the length of a spring when different weights were hung on it. They found 'essentially no correlation for an individual student's scores. Students with either a 9 or a 1 Spring score had Paper Towels scores ranging from 1 to 9.'[86] Since a particular task is one from the wide range of possible tasks, the 'task sampling error', is large and it means that to obtain a reliable score for an individual student would require the individual to tackle a totally unacceptable number of tasks. The only meaningful way of reporting performance is to describe each investigation and the range of different ways in which students respond.

Alternatives to tests

It appears from this discussion that attempts to provide dependable assessment of understanding and inquiry skills using tests come up against some unavoidable obstacles. We noted earlier some key requirements for assessing both scientific understanding and science inquiry skills. These were

- that students are actually involved in undertaking inquiry
- that they are engaged with a question that is novel to them
- that the knowledge which is inevitably required, but is not being assessed, is available to them.

When students are presented with a task in a specified context and confined by a particular format, as in a test, there is bound to be some uncertainty about meeting these requirements. Genuine inquiry takes place when students seek to answer a question that is new to them and to which they do not already know the answer. But who is to judge what is 'new' for a particular student? Can the response of a student created in isolation from the normal context of learning and interaction with others really reflect their capability?

Research conducted in Denmark[87] using items from the PISA tests provide a clear answer to this last question. The research involved students in answering some PISA questions orally in an interview and conducting, in pairs, an investigation described in a PISA item. The conclusion reached was that 'when compared directly and following the scoring criteria of PISA, pupils' performance increased by 25% when they were allowed to exercise their knowledge in a socio-culturally oriented test format.'

There are many other questions and problems that point to the deficiencies of tests in addition to the issues arising when the results are used for high stakes decisions and judgements as noted in Chapter 4.

Here we turn to the question: are there any alternatives to tests for summative assessment? Fortunately there are; all of them depending on the fact that the experiences that students need in order to develop desired skills, understanding and attitudes also provide opportunities for their progress to be assessed. The key factor is judgement by the teacher. Assessment by teachers can use evidence from regular activities supplemented, if necessary, by evidence from specially devised tasks introduced to provide opportunities for students to use the skills and understanding to be assessed.

[86] Pine, J., Aschbacher, P., Rother, E., Jones, M., McPhee. C., Martin, C., Phelps, S., Kyle, T. and Foley, B. (2006) Fifth graders' science inquiry abilities: a comparative study of students in hands-on and textbook curricula, *Journal of Research in Science Teaching* 43 (5): 467-484. P 480

[87] Dolin, J., & Krogh, L. B. (2010): The Relevance and Consequences of Pisa Science in a Danish Context. International Journal of Science and Mathematics Education, 8, 565-592.

Over the period of time – such as a semester or half year – for which achievement is being reported, students have opportunities to engage in a number of activities in which a range of competences can be developed. These same activities provide opportunities for this development to be assessed by the teacher. In other words, the limitation on the range of evidence that can be obtained through a test does not apply when assessment is teacher-based.

There are other advantages that go beyond more valid assessment of understanding and inquiry skills, since a greater range of competences can be included. Observation during regular work enables information to be gathered about processes of learning rather than only about products. Even the kinds of items shown in Example 5 are not capable of assessing qualities such as reflection on the process of learning. Such information is useful to those selecting students for advanced vocational or academic courses of study, where the extent to which students have learned how to learn, and are likely to benefit from further study, is as important as what they have already learned.

Implementing teacher-based summative assessment

Summative assessment by teachers involves deciding, collecting, interpreting and communicating evidence to provide a summary of students' achievement. Just as there are different types of task and ways of responding to them in the case of tests, so there are different approaches to assessment by teachers according to how, when and where evidence is collected and how it is interpreted and used. Before looking at some these we should make clear that assessment by teachers is not just a matter of teachers using their individual judgements about what evidence to use and how to interpret it. Summative assessment by teachers uses agreed procedures and is subject to quality control measures appropriate to the use of the results, that is, stricter control for higher stakes use.

The process requires: some *evidence or data*, *criteria* by which to judge it and *procedures* for arriving at a judgement.

Evidence or data relating to science understanding and inquiry skills may come from

- observing students involved in scientific investigation
- a portfolio of work collected over a period of time including accounts, reflections, photographs and other products of inquiry
- students' note-books and /or electronic postings
- presentations made by students individually or in groups.

Criteria vary according to the degree of specification. For example, a highly specified criterion would be *'the student knows that magnets attract certain materials but not others, and can repel each other'*. This can only be applied to work relating to magnets. Whereas *'the student can group materials according to their physical properties'* can be applied to information from a variety of learning tasks.

A high level of specification of criteria means that teachers are required to base their assessment on certain pieces of work. This occurs in some forms of portfolio assessment, where types of task to be included in the portfolio are closely prescribed and criteria given for each type. However, the more the tasks are specified the less opportunity there is for evidence to be collected from a broad range of activities, which is at the heart of reasons for using assessment by teachers. A review of research on the validity and reliability of assessment by teachers[88] concluded that the most dependable approaches were those where criteria were detailed but generic, being applicable to a range of classroom activities. Research indicates that well specified criteria help teachers to make reliable judgements. The most effective criteria guide the selection of evidence without prescribing it.

[88] Harlen, W. (2004) Trusting teachers' judgements: research evidence of the reliability and validity of teachers' assessment for summative purposes, *Research Papers in Education*, 20(3); 245-270.

Procedures for making judgements may start from the criteria, then search for evidence that meets them, or start from the evidence and then search the criteria to see which if any best describe the work. In most cases there is a combination of these approaches, with a to-ing and fro-ing between the data and criteria in order to find the 'best fit'. Rather than trying to match a particular piece of work to a particular criterion, teachers take evidence from several relevant activities and form an 'on balance' judgement as to the criterion that best matches the evidence. Criteria also serve the additional function of focusing attention on the outcomes of particular kinds of work so that teachers, alerted to looking for particular behaviours, are less likely to miss them.

It is common for criteria to be identified at different 'levels', as in Box 19 (page 64), so that the outcome of the assessment can be expressed in terms of the level at which a student is currently performing. Levels are produced by mapping the progress of students in a particular area of learning, using what research evidence there is together with evidence from teachers' experience and some trial and error. Examples are the Developmental Assessment materials, developed in Australia[89] and the descriptions of performance at various levels set out in a number of national assessment standards and targets, such as the National Curriculum for England (Box 19). However, experience of use (or, more accurately, misuse) of 'levels' has led to some reaction against their use in reporting on students' performance. The pros and cons of using levels is discussed later (page 69) since the arguments apply equally to using tests. We now turn to some ways of improving the reliability of assessment by teachers.

Improving the reliability of teacher-based assessment

The most commonly expressed criticism of assessment by teachers concerns the reliability of the results. It can indeed be the case that, when no steps are taken to assure quality, teachers' judgements are prone to a number of potential errors. Research studies have reported evidence of bias – a non-random source of error – due to teachers taking into account information about non-relevant aspects of students' behaviour or being apparently influenced by gender, special educational needs, or the general or verbal ability of a student, in judging performance in particular tasks. For example, general behaviour was found to influence teachers' judgements of younger students' achievements. As classroom behaviour is often linked to gender in the early years of school (little girls tend to be better behaved than little boys), the result is sometimes reported as gender bias. The effect is much less for older students. Other causes of low reliability arise from the inclusion of irrelevant information (such as neatness or spelling, when these are not specific goals of the task), variation in interpretation of the criteria, and the problem of relating performance in specific contexts to necessarily more general criteria. However, there are several effective ways in which reliability can be improved, to a level equal to and even exceeding that of tests. The main ones are group moderation, using examples, and using a test or tasks as a check.

Group moderation

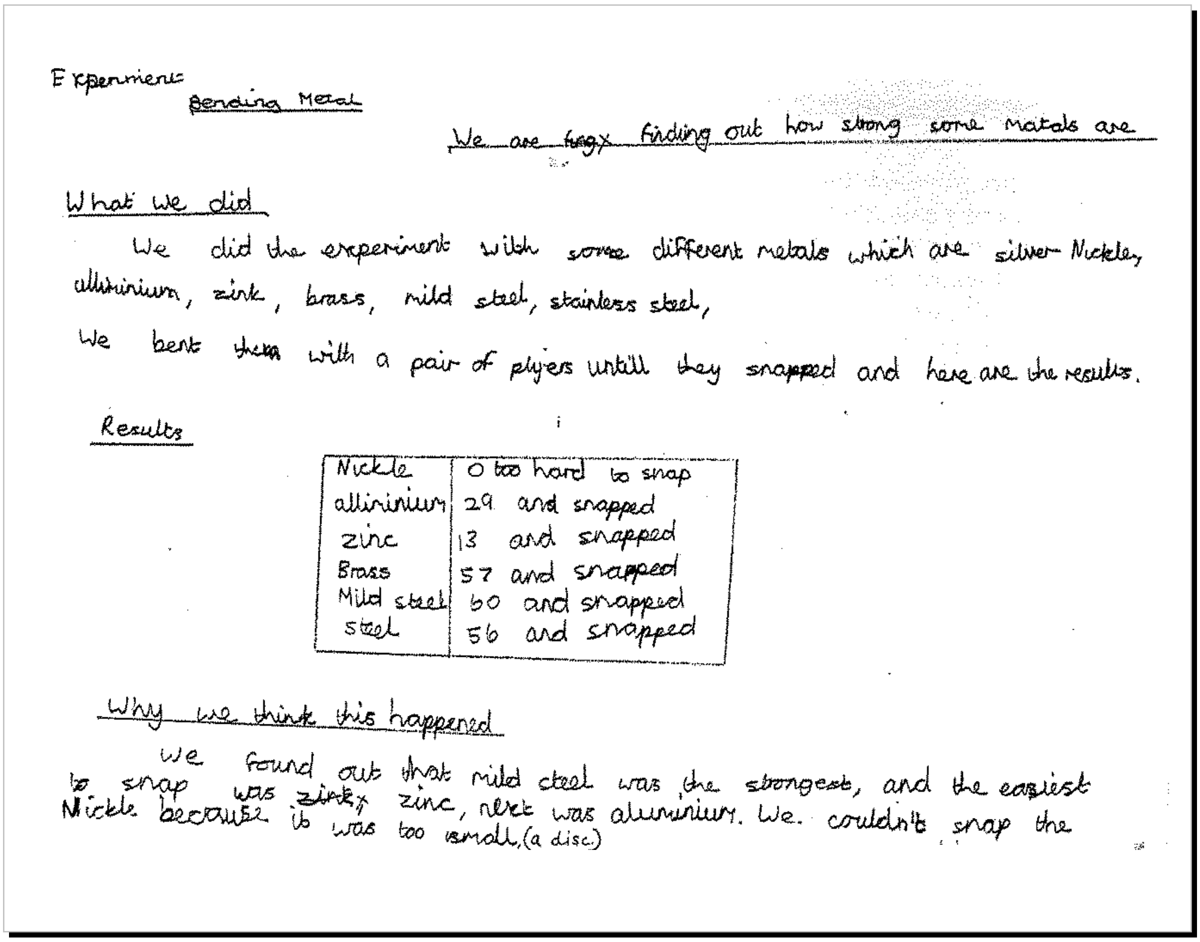
This involves teachers meeting to review samples of students' work, but the purpose is not to verify decisions about particular students' work, rather to arrive at shared understandings of criteria and how they are applied. The intention is to influence the process of assessment and in this way ensure greater reliability of the results for all students. Teachers bring to the meeting several examples of a student's work, since judgements should not be made on the basis of single piece of work, describing the context of the work and discussing with other how judgements were made. Group moderation has benefits beyond improving the quality of assessment results. It has well established professional development functions. Meeting to discuss the inferences that can be drawn from observing students and studying their work provides teachers with insights into the assessment process and improves not only their summative assessment but also their formative use of assessment.

[89] Masters, G. and Forster, M. (1996) *Progress Maps*. Camberwell, Victoria, Australia: ACER

Using examples

Providing examples of students' work and showing how certain aspects relate to the criteria of assessment helps in conveying what the criteria mean in practice. Good examples also indicate the opportunities that students need in order to show their achievement of skills or understanding. Although the outcome of assessment is identified in the examples, the focus is on the process of arriving at that outcome. There are many sources of examples of students' work or descriptions of their actions, annotated to highlight features which are significant in relation to the judgements to be made. Some curriculum materials contain examples of students' actions, words, talk, writing or drawings and discuss aspects which lead to a decision about whether certain key criteria have been met. For example the Nuffield Primary Science (1995) Teachers' Guides each include a chapter on assessment, where students' work is reproduced with a commentary provided on the aspects that led to a judgement of the level reached. Example 6 is taken from the Teachers' Guide on *Materials* for students aged 8 to 11 where one student writes about his group's investigation of the hardness of different metals. (This one piece of work is used for illustration; in practice a single piece of work alone should not be used for assessing the level of performance.)

Example 6: Matthew's report of his group of 10-year olds' investigation of the strength of metals



Experiment
Bending Metal

We are ~~fing~~ finding out how strong some matals are

What we did

We did the experiment with some different metals which are silver Nickle, alliminium, zink, brass, mild steel, stainless steel,
We bent them with a pair of plyers untill they snapped and here are the results.

Results

Nickle	0 too hard to snap
alliminium	29 and snapped
zinc	13 and snapped
Brass	57 and snapped
Mild steel	60 and snapped
steel	56 and snapped

Why we think this happened

We found out that mild steel was the strongest, and the easiest to snap was ~~zink~~ zinc, next was aluminium. We couldn't snap the Nickle because it was too small, (a disc)

Box 19: Example of assessment criteria relating to scientific inquiry

(Note: one level spans approximately two years, level 2 being the expected level for students aged about 7)

Level 2: Making suggestions as well as responding to others' suggestions about how to find things out or compare materials. Using equipment to make observations. Recording what they find and comparing it with what was expected.

Level 3: Saying what they expect and suggesting ways of collecting information to test their prediction. Carrying out fair tests and knowing why they are fair. Recording what they find in a variety of ways. Noticing any patterns in their findings.

Level 4: Making predictions which guide the planning of their inquiry. Using suitable equipment and making adequate and relevant observations. Interpreting, drawing conclusions and attempting to relate findings to scientific knowledge.

Level 5: Planning controlled investigations of predictions which are based on scientific knowledge. Using graphs, charts or tables to record and help interpretation. Considering findings in relation to scientific knowledge.

The criteria used in the assessment (Box 19) were derived from the 1995 English national assessment attainment targets for scientific inquiry.

The commentary on example 6 explains:

> Matthew's group's investigation followed from their claim that metals were strong. The teacher encouraged them to find out more about this and to find a way of seeing whether different metals were as strong as each other. Their investigation was to compare how easily certain metals could be snapped when repeatedly bent forward and backward using a pair of pliers. They quantified their observations by counting, but made no mention of the need to control the amount of bending in each case, nor the size of each piece of metal. However, they made their observations systematically, using a table and they interpreted their finding, indicating work approaching level 3. More attention to fairness of their tests is needed for evidence of achievement at level 3.[90]

Since teachers should be basing their judgements on a range of each student's work and not judging just from one piece, it is most useful to have exemplar material in the form of a portfolio of work from one student than single pieces of work from several students. This helps teachers to apply the criteria in a holistic manner. A glance at the level descriptions for inquiry skills in Box 19, for instance, shows that not every piece of work will fit the descriptions and neither will each and every part of the criteria for a level be represented in the portfolio. The approach recommended is that of 'best fit', comparing the evidence with the most likely level and with those just below and above it.

Examples can be employed in group moderation, but are particularly useful for individual teachers unable to participate in group moderation meetings.

[90] Nuffield Primary Science Teachers' Guide *Materials*. (1995) London: Collins Educational. p 86

Using a test or task

In this approach a common short test or special task is used as a means of moderation or checking teachers' judgements but not as a separate measure of achievement. Tests are used in this way in Scotland for assessing English and mathematics. Teachers decide on the basis of a range of evidence from everyday activities conducted over time about whether a student has met the criteria at a particular level in the subject. One option for moderating this judgement is for teachers to use a short test at the level indicated by the teachers' own assessment. The test, which teachers administer and mark themselves, is drawn from an externally devised bank of items. The teacher compares the test results with the results of their own classroom assessment. This can be done whenever the teacher judges that a student can pass the test, not at a fixed time, and is administered on an individual basis in an informal manner that prevents test anxiety in students. When confirmed by moderation, the level reached according to the teachers' judgement is recorded and then reported at the appropriate time.

Examples of teacher-based assessment

Teachers' assessment of students is not new, of course; it is commonly used for routine within-school reporting and record-keeping where there are no high stakes attached to the results. In such cases less rigorous moderation is needed than in cases where important decisions may be made on the basis of the results. An example of assessment by teachers being used for assessment which has high stakes for students has already been mentioned in Chapter 3. Some further points are relevant here.

End of secondary school teacher-based assessment

The operation of the system of school-based assessment for the Senior Certificate in Queensland, Australia, is a well-documented example of how such a system can work. It has been in existence since 1972 when Queensland abolished external examinations. The reasons for the preference for assessment by teachers reflect the arguments that such assessment is able to include a range of learning outcomes, both academic and vocational and to support rather than to control the curriculum. In addition, Maxwell argues that

> An important principle of school-based assessment is that the assessment is progressive and continuous. One of the aims…is to alleviate the peak pressure of a single final examination – the one-shot test on which everything depends. This requires not only that the assessment is tailored to the way in which each subject syllabus is implemented by the school but also that assessment occurs progressively over the whole course of study. In other words, the validity of the assessment is improved by assembling the portfolio from a variety of assessment types and contexts. So, too, is the reliability improved by having many opportunities for the student to demonstrate their knowledge and capability and by collecting the information on many different occasions.[91]

The process is portfolio-based and allows for variation in the content so that syllabuses can be implemented with flexibility to meet local requirements. The common element is the system of progressive criteria against which each portfolio is judged. There is also a strong system of moderation in the case of those subjects that count towards university entrance, which successfully assures confidence of all concerned in the outcome of the assessment.

The portfolio is built up over the two years of the course, during which time its content will change not only through addition of new material but through replacing older by more recent evidence. It is only the final evidence that is taken into account, although some will have been collected earlier than other.

[91] Maxwell, G. (2004) op cit. p 2

Thus the final or 'exit' portfolio is made up of assessment tasks that represent the *fullest and latest* information of the student's knowledge and capability. As Maxwell explains,

> *Fullest* information means that assessment information must be available on all mandatory aspects of the syllabus. Important criteria cannot be skipped; the assessment evidence in the portfolio must cover all the required aspects of the course…*Latest* information means that earlier assessments that are no longer relevant may be discarded and replaced by more recent evidence… The ultimate aim is to represent the state of knowledge and capability as typically demonstrated by the student towards the end of the course.[92]

The criteria for assessment, which are published so that students and parents as well as teachers can be familiar with them, describe what students do in terms of grade descriptions. For example, one of the sub-categories of 'working scientifically' relates to planning and the criteria for this are set out at five levels or standards from A downwards in Box 20.

Box 20: Criteria for grading portfolios relating to planning (Queensland Senior Certificate)

Standard A: plans a range of scientific investigations of problems including many with elements of novelty and/or complexity

Standard B: plans a range of scientific investigations of problems including many with elements of novelty and/or complexity

Standard C: plans a range of scientific investigations of straightforward problems

Standard D: participates in planning some scientific investigations of straightforward problems

Standard E: participates in some aspects of planning scientific investigations of straightforward problems

In Box 20 the criteria for standards A and B are the same, but the judgement of 'planning' is only one of several aspects of 'working scientifically' to be judged as a whole. The comparison of evidence with criteria involves judgements, not aggregation, and is an 'on-balance' judgement of best fit.

Moderation involves several stages, beginning with approval of the school's 'work plan'- the details of how the school intends to provide opportunities for the student to meet the final criteria for assessment in a subject. Moderation of those subjects counting towards university selection involves external district panels who review sample portfolios from schools and consider evidence that supports or challenges the schools' judgements. In turn a state panel reviews samples from districts and arbitrates difficult cases.

The openness of the on-going process of creating the portfolio means that at the end of the course there should be no surprises for either teachers or students. Further, the 'selective updating' and collection of 'fullest and latest' evidence allow poor starts, atypical performances, and earlier and temporary confusions (for whatever reason) to be ignored. Importantly, these processes facilitate the use of assessment to help learning, for students benefit from the feedback they receive on earlier assessments. They also have the opportunity for self-assessment in deciding when to replace an earlier piece of work in their portfolio.

The Queensland experience supports the value of collaborative moderation not only in assuring reliable assessment but in providing professional development.

> The most powerful means for developing profession competence in assessment is the establishment of regular professional conversations among teachers about student performance (moderation conversations). This is best focussed on actual examples of

[92] Maxwell, G. (2004) op cit. p 4-5

> student portfolios. Looking at actual examples and discussing the conclusions that can be drawn about the student's performance when compared to explicit standards sharpens teachers' judgement and builds knowledge and expertise about assessment more successfully than any other process.[93]

End of primary school teacher-based assessment in England

An example of teacher assessment of younger students is provided by the introduction in England in 1989 of assessment by teachers as part of the national assessment of students at ages 7, 11, 14 and 16 in English, mathematics and science. As well as reporting their own assessment of performance levels achieved by their students, teachers were required to administer externally designed tests. When first introduced[94] it was recognised that the external tests could cover only parts of the curriculum; teachers' assessment was intended to be the major component. (This recognition did not last, for reasons noted in Chapter 4). Formal reporting of their assessment against national curriculum levels was novel and attracted the attention of researchers to find out how primary school teachers went about it. Gipps et al[95] observed and interviewed teachers of students aged 11. They found several different approaches but most had in common the day-to-day collection of data by using techniques such as standing back, listening, asking open-ended questions, observing and either making notes or using their memory. Most teachers aimed to disturb normal teaching as little as possible. Some kept significant pieces of students' work; others kept a record of the marks given to students' work. If teachers found they did not have enough data for some students in relation to the aims of the work, they would target these students for attention. At the end of the year, when it was necessary to assign a level to each student's work, teachers used the work collected and/or their notes and recollections to decide the level of achievement using the 'best fit' approach against level-based criteria similar to those in Box 19.

Teacher-based assessment of inquiry skills in primary schools in France

A significant change in the curriculum was introduced in France in 2006 in the form of *Le socle commun de connaissances et de compétences* (Common Base of Knowledge and Skills). This prescribes what knowledge and skills students must have acquired by the end of compulsory education (age 16) and gives to the schools the responsibility for achieving this. It was, and remains, a very great change in schools' responsibilities. To implement the part of this new curriculum relating to science, primary and middle school teachers were faced with the considerable challenge of teaching and assessing science inquiry skills, which most had not done before. The Ministry provided a record booklet for each student, in which teachers are intended to record, at the end of each year, whether or not the student has achieved specified skills and knowledge, as a simple 'yes' or 'no'. To provide help to teachers with the implementation and the assessment of inquiry skills the project team of *La main à la pâte* (LAMAP) helped local groups to develop tasks for students that would enable teachers to assess inquiry skills and knowledge. Tasks included both closed and open questions and some practical activity, as in Example 7.

In addition, the Ministry provided a guide for teachers to observing and questioning students to help them to assess their students both formatively and summatively.[96] Four sources of assessment data are proposed: the study of students' notebooks; observation during class activities; students' written or oral accounts of a practical investigation, including what they did and what they found, and illustrated by diagrams; giving a standardised test. Guidance is also given as to what to look for when reviewing students' written work, when observing them in class and when listening to their presentations.

[93] Maxwell, G. (2004) op cit. p 7

[94] DES/WO (1988) *National curriculum Task Group on Assessment and Testing: A Report.* London: HMSO

[95] Gipps, C., McCallum, B. and Brown, M. (1996). Models of teacher assessment among primary school teachers in England., *The Curriculum Journal*, 7 (2) 167-183

[96] http://cache.media.eduscol.education.fr/file/socle_commun/99/7/Socle-Grilles-de-reference-palier2_166997.pdf, page 36

Example 7: A task for end of primary students in France from LAMAP[97]

In this extended task, students first read a page of information about ice cover around the north and south poles and how it changes from winter to summer in both regions. They answer some questions on the information given including the effect of rise in temperature on the ice at the poles. An investigation is then suggested and illustrated by a photograph showing two glasses of water. One glass shows an ice cube floating in the water; the other shows an ice cube held above the surface of the water. The level of the water in both glasses is the same. Students are asked to predict what will happen to the ice and to the level of water in the two glasses after a period of time. They have to give reasons for their answers. Students are intended to conduct this investigation in practice, but there is also another photograph showing what happened after the ice melted in case this is not possible. A series of questions asks students to say whether what happened agreed with their prediction and to assess the evidence for a series of possible reasons for the observations. A final question asks students to comment on the statement that is often made about global warming that 'the melting of polar ice will cause the level of the sea to rise'. An extended written answer is requested using appropriate vocabulary.

Dual use of data for formative and summative assessment

In these three examples it is seen that there are opportunities for the formative use of the data being collected for summative assessment. This dual use is more conscious and deliberate in the following (admittedly idealised) account of a series of science lessons in which year 8 students (aged 13) were studying the transfer of heat energy through different materials.

> At one point the students were investigating the insulating properties of various materials that could be used to make coats. They were provided with metal containers in which they could put water to represent a 'body' and pieces of fabric to wrap around the outside of the containers. Thermometers were placed in the water to measure changes in the 'body' temperature. But there were several decisions to make in setting up the investigations. What should the temperature of the water inside the container be? Would it give a useful result to carry out the investigation in the warm laboratory instead of the cold outside? How to make sure that the comparison was fair? Some of these decisions required the students to apply what they knew about conduction and other ways of transferring heat energy whilst others required understanding of how to make fair comparisons.
>
> While they were planning what they would do the teacher drew their attention to the list of things to consider in planning an investigation, a list which had been developed from an earlier discussion and which was in their folders. The teacher observed their actions and listened to their talk as they planned and carried out their investigations, occasionally asking groups for explanations of their reasons for how they doing certain things. At the stage of reporting on their findings there was further opportunity for the teachers to gather evidence that could be used to help in developing the students' understanding of how heat energy is transferred and their enquiry skills.
>
> Thus, during the lesson, the teacher responded to what she saw and heard by raising questions to provoke rethinking of decisions; asking for justification of statements or actions; asking for explanations of how certain parts of the investigation would help them achieve their goal. In other words, the teacher was using evidence formatively to help learning. She also made notes of the help that some students had required so that this could be followed up in future lessons.

[97] http://www.fondation-lamap.org/fr/page/14209/des-exemples-d-valuation-adapt-s-des-s-quences-d-enseignement

> Then, at the end of the year, when it was necessary to report progress in terms of levels achieved, the teacher reviewed the evidence from this and other science lessons. For both the conceptual goals and the inquiry skills goals, evidence from different specific activities had to be brought together so as to form an overall judgement about each student's achievement. In preparation for this the teacher made time available for the students to look through their folders, compare their later accounts of investigations with their earlier ones in terms of how evidence had been collected and used to reach a result and an explanation of the result. They then picked out the best examples of work with these features. By providing lesson time for this the teacher was able to talk with individuals to ensure that the selection criteria were understood and appropriately applied. She then reviewed the evidence against the criteria for reporting levels of attainment, in this case descriptions for levels 4, 5 and 6 of 'scientific enquiry' and of 'physical processes' in the English national curriculum. The results were reviewed by the head of department and the evidence for a sample of three students was discussed at a departmental moderation meeting.[98]

Reporting summative assessment: pros and cons of using levels

Levels, as noted earlier, set (national) standards against which students' achievement is judged. Standards embody a view of progress in students' learning in descriptive terms but can also be converted into a scale, in which levels are defined in terms of points in the overall progress. Levels are therefore a short-hand way of signalling learning achieved and have the advantage of enabling analysis of results of groups of students. Test scores can also be converted into levels by assigning levels to particular ranges of scores.

Problems with levels

However, the use of levels has been challenged in the course of curriculum and assessment revision in England for reasons which no doubt have relevance in other systems where numbers of students reaching certain levels has become a high stakes measure of teacher and school. The report of an influential expert group in 2012[99] made some strong criticism of the use of levels for reporting students' progress, particularly in the primary and middle school, accepting that students are inevitably differentiated in the later years of secondary school. The main points, which may resonate in other assessment systems, are:

- The award of 'levels' encourages a process of differentiating learners to the extent that students come to label themselves in these terms.
- Some students become more concerned for 'what level they are' than for the substance of what they know, can do and understand.
- Assigning levels increases social differentiation rather than striving for secure learning of all students.
- Describing a student as having achieved a certain level does not convey anything about what this means the student can do, nor does it indicate what is necessary to make progress.
- When levels are seen as important, teachers, parents and students use them inappropriately to label students.
- Students who are regarded as unable to reach target levels often have reduced opportunities for progress, increasing the performance gap between the more and less well achieving students.

[98] Harlen, W. (2007) Op cit p129

[99] Department for Education, (2011). *The Framework for the National Curriculum. A report by the Expert Panel for the National Curriculum review.* (London: Department for Education).

- As level are generally well spaced (the levels being about two years apart) the practice has grown of creating sub-levels in order to be able to show progress. However, these have little basis of evidence in cognitive development and serve only to prescribe the curriculum more closely.

Overall, then, reporting students' performance, whether assessed by test or by teachers' judgement, as a 'level' of achievement has been found to have a profound impact on how teachers, parents and students themselves judge progress, with implications for pupil motivation and learning. These are unintended consequence of an over-prescriptive framework for curriculum and assessment.

Alternatives to levels

An alternative to reporting by levels is to identify learning outcomes to be achieved by the end of key stages of education. Key stages might span two or 3 years, so the years from Grade 1 to 12 might be divided into five or six stages. The expected learning outcomes might well be similar to the statements in curriculum progress maps, but focusing on a few key competences. Judgements would be about the achievement of these key competences with the explicit aim of ensuring that they are achieved by all students by the end of the key stage. Teachers' judgements during the key stage would identify where students are on course to achieve the expected outcomes or where they need special attention. Thus the assessment would change from seeking the 'best fit' on the ladder of levels to tracking the achievement of key competences.

The experience of those countries with high levels of achievement and small gaps between the higher and lower attaining students, does not suggest that the consequence of this approach would be students repeating years. What characterises high-performing systems appears to be their approach to student progression and differentiation. Instead of crude categorisation of students' attainment there is encouragement for all pupils to achieve adequate understanding before moving on to the next topic or area. There is no assumption that ability limits achievement, which can have a negative effect on expectations of students. Rather it is assumed that deep engagement and effort leads to understanding and achievement for everyone.

Advantage and disadvantages of tests and teacher-based assessment

In summary we bring together some of the advantage and disadvantages of using tests or teachers' assessment for summative purposes. The main factors concern the trade-off between validity and reliability (see Chapter 1 p10). Tests provide information about only a sample of goals but steps can be taken to ensure optimum reliability. Striving for high reliability, however, leads to the inclusion of goals most easily tested reliably, thus tending to exclude less easily tested goals, such as higher level skills and practices. In the case of assessment by teachers, a wider range of achievement and learning outcomes can be included but reliability may be low unless steps are taken to ensure that comparable standards are being applied. However, moderation procedures provide valuable professional development for teachers.

Other factors to take into account in deciding the balance of advantage in a particular case are:

- Tests provide teachers with clear examples of the meaning of learning goals, but they direct teaching in specific directions which are the same for all students. Assessment by teachers allows teachers greater freedom to pursue learning goals in ways that suit their students.
- Tests that are provided externally to the school enable teachers to distinguish their role as teacher from that as assessor. Responsibility for assessment unites the role of teacher and assessor and may seem to increase teachers' workload.
- In some circumstances the results of tests can be used to provide feedback that helps learning, but generally these opportunities for formative use are limited. When teachers gather evidence from

students' on-going work, it can be used formatively, to help learning, as well as for summative purposes.

- It is well known that tests induce anxiety in many students. Not all are affected equally and the impact is often exacerbated by frequent test-taking practice. Assessment by teachers reduced this source of error and inequality.
- Time spent by teachers in preparing students for tests can be more effectively used for learning when assessment is on-going. Financial resources are also released when fewer commercial tests are purchased.
- Users of assessment data often have more confidence in tests than in teachers' assessment, especially for older students. Any change requires greater openness about both test accuracy and procedures that can enhance assessment by teachers.

Chapter 7
Changing assessment practices

The interaction between assessment, curriculum content and pedagogy represented in Figure 3 (page 26) signals that change in assessment will involve wider change in practice. It will not be just a matter of adding new procedures to otherwise unchanged classroom interactions and processes. This is particularly the case when the changes in assessment concern greater use of teachers' judgements, requiring skills and understanding that are different from traditional teaching and testing. Although there is as yet limited experience of change in assessment of IBSE, beyond the scale of pilot trials, there is much to be learned from how change in other aspects of education has been brought about. Thus in this chapter we first look briefly at the lessons that can be learned from attempts to make change in curriculum content and pedagogy. A review of some examples of making change in assessment then leads to some suggestions of what may be needed in relation to change in assessment of IBSE. Whilst the experience described here has been in the context of professional development of serving teachers, the approaches and principles are equally relevant to initial teacher education.

Approaches to changing practices in education

Different approaches to changing practices in education can be divided into two main groups: transmission and transformation.[100] Various forms of *transmission* involve the distribution of resources giving ideas and examples of new content and practices after their development and publication. The resources produced are often written guides and associated audio-visual materials. When they have been created to match new curricula they give the promise of a neat solution to the need to make required changes. It seems that all that teachers are required to do is to follow the guides, since the necessary thinking will have been done for them. This 'top down' approach has, however, fallen out of favour through recognition that the messages received and acted upon in the classroom rarely match the intentions of the producers.

The evident need for some mediation of the messages through professional development sessions led to the 'cascade' approach. This starts with a group being trained in the subject and the intended practices and sometimes in ways of training other groups. Those trained then train others and the training passes from group to group and eventually to classroom practitioners. This approach has well known disadvantages: there is much potential for the distortion of the messages at each stage of the cascade. 'Pilot and roll out' is another approaches with some similarities, but starting from trial and modification of ideas in a few schools to ensure practicability and provide credibility among teachers.

The limited success of transmission approaches has led to the alternatives described as *transformational*. Experience in many areas of change in education, be it in the curriculum, pedagogy, assessment or school organization and management, is that participation in developing new procedures or materials by those who have to implement them is a most effective way of encouraging commitment to change. When groups of teachers work together with researchers or developers they can be creative and experimental in a safe environment, learn from each other, combine ideas, and achieve ownership of the emerging practices in what is described as a 'bottom up' approach. Add opportunities to reflect and develop understanding of principles underlying the change, and the

[100] Hayward, L. (2010) Moving beyond the classroom, in J. Gardner et al *Developing Teacher Assessment*. Maidenhead, UK: Open University Press.

experience can be a most effective form of professional learning. But it is also very demanding of resources, particularly time, and clearly cannot easily be extended to large numbers of teachers. It does, however, indicate the kinds of experiences that are likely to be effective in cases where change requires more than adopting new techniques.

Transformational approaches recognise that change in practice is a matter of learning and that effective learning by teachers has the same qualities as for students. Just as students develop understanding through their own mental and physical activity, so teachers learn best when they take an active part in transforming their practice. Hayward[101] quotes Bruner[102] in arguing for four essential components of effective learning:

- agency, individuals taking more control of their own learning activity
- reflection, making sense of what is being learned, understanding it, internalizing it
- collaboration, sharing the resources of the 'mix of human beings' involved in teaching and learning
- culture, the construction by individuals and groups of particular ways of life and thought that they call reality.

A further factor, revealed in evidence from three large scale development projects in the Danish secondary school,[103] is the extent to which an innovation is seen by teachers as being consistent with their (often tacit) beliefs and values in education. When there is a large degree of agreement teachers are more likely to embrace new ideas and make necessary changes in their practices. If there is some conflict with their values, teacher participation may well lack real engagement and the project have little influence on their practices.

In transformational approaches, rather than assuming that all classrooms are the same – and thus that one solution will fit all – the assumption is that different learning environments will require different solutions. Of course the problem of reaching large numbers of teachers remains. The opportunities to involve teachers in genuine development and to tailor experiences to individual needs on a large scale are clearly limited. However, it is possible to provide situations in which teachers can learn collaboratively and work out how to put new ideas into practice and achieve new goals in the particular context of their own classrooms as some of the examples later show. The opportunity for teachers to visit others' classrooms or to show videos of their teaching has been found effective in enhancing teachers' learning, while projects such as Fibonacci[104] and Pollen[105] have shown what can be done when time and resources are used to bring teachers and teacher educators together to learn from each other.

Some examples of changing practices in assessment

Both the practices of using assessment to help learning and of summative assessment in which teachers take a leading role require strategies that differ considerably from those familiar to many teachers. The early work in changing assessment practices focused on formative assessment, once it became widely known, through the work of Black and Wiliam,[106] that implementing aspects of formative assessment practice could significantly raise the level of students' achievement (see Chapter

[101] Hayward, L (2010) op cit p 96

[102] Bruner, J. (1996) The Culture of Education. Cambridge, MA: Harvard University Press

[103] Dolin, J., Laursen, E., Raae, P. H., Senger, U. (2005). Udviklingsprojekter som læringsrum. Potentialer og barrierer for skoleudvikling i det almene gymnasium. Gymnasiepædagogik nr. 54, Syddansk Universitet. 232 s. (Development projects as learning arenas. Potentials and barriers for school development in the general upper secondary school. University of Southern Denmark)

[104] www.fibonacci-project.eu

[105] www.pollen-europa.net

[106] Black, P. and Wiliam, D. (1998) Assessment and classroom learning, *Assessment in Education*, 5 (1) 7-74

3). For many, this was a novel use of assessment and one which achieved noticeable positive responses from students. At the same time, it was realised that existing approaches to summative assessment did not provide information in line with the goals of a modern education and that to improve this matching requires a greater involvement of teachers in the process.

Teachers developing formative assessment practices: experience in England

The King's Medway Oxford Formative Assessment Project (KMOFAP) was the forerunner of several projects developing the formative use of assessment.[107] KMOFAP started soon after Black and Wiliam completed their review of classroom research and recognised the importance of finding ways in which teachers could incorporate formative assessment into their work. University researchers from King's College and advisory staff from two local authorities collaborated in planning work with two mathematics and two science teachers from each of six secondary schools. Funding enabled teachers to be released for professional development sessions comprising seven whole day sessions spread over 18 months. After introducing teachers and advisers to the aspects of practice that the research showed were effective in improving learning, the aim of the first few sessions was for teachers to draw up action plans for implementation in the subsequent school year. This gave teachers time to 'experiment with some of the strategies and techniques suggested by the research, such as rich questioning, comment-only marking, sharing criteria with learners, and student self -assessment and peer - assessment'.[108]

During the project, team members visited teachers' classrooms to observe and discuss what was happening and how it related to the action plans. These visits were described as being 'not directive, but more like holding up a mirror to the teachers'.[109] The teachers were introduced to general strategies and some idea of what to aim for, but not given models of what to do. The input from the researchers was thus designed to engage teachers in identifying how to put various features of formative assessment into practice. This was an unusual and perhaps uncomfortable role for the teachers to be given by researchers and advisers whom they regarded as experts. The researchers noted that

> At first, it seems likely that the teachers did not believe this. They seemed to believe that the researchers were operating with a perverted model of discovery learning in which the researchers knew full well what they wanted the teachers to do, but didn't tell them, because they wanted the teachers 'to discover it for themselves'. However, after a while, it became clear that there was no prescribed model of effective classroom actions, and each teacher would need to find their own way of implementing these general principles in their own classrooms.[110]

In the course of trying out the activities in practice, the teachers became aware of the need to understand why these particular activities are important and why they 'work'. As teachers used the activities and saw the reaction of their students to them, they wanted to know more about the way students learn. So, approximately a year after starting the work focusing on classroom actions, one of the in-service sessions was designed to introduce teachers to theories of learning and explain the importance of students taking an active part in their learning.

[107] This account draws on Harlen, W. (2010) On the relationship between assessment for formative and summative purposes, in Gardner, J. et al *Developing Teacher Assessment* Maidenhead, England: Open University Press pages 100 – 129.

[108] Wiliam, D. Lee, C., Harrison, C. and Black, P. (2004) Teachers developing assessment for learning: impact on student achievement, *Assessment in Education*, 11 (1) 49-66, p 54.

[109] ibid, p 54

[110] Ibid, p 51

In terms of a strategy for transforming research findings into classroom practices, the researchers reported that the new practices developed by teachers were

> far more rich and more extensive than those we were able to suggest at the outset of the project on the basis of the research literature. All of them involve(d) change in the way that they work with their students and with the curriculum.[111]

In terms of change in teachers, there were the expected differences in response among the 24 teachers. However, the researchers reported changes in several respects in all the teachers involved in the project. In particular they noted change in:

- the way teachers thought of their goal as being to help students learn, in contrast to 'getting through the curriculum'
- the teachers' expectations of their students, in that all were able to learn given time and the right approach
- teachers giving students more control and sharing with them the responsibility for learning.

Black and his colleagues claimed that the changes in practice were slow to appear but were lasting, and that they were unlikely to have occurred had the researchers provided recipes for successful lessons. They considered that 'the change in beliefs and values are the result of the teachers casting themselves as learners and working with us to learn more'.[112] Wiliam et al,[113] using a variety of comparison classes and internal and external school examinations and national curriculum test results, reported a positive impact on students' achievement even after one year of intervention.

Scaling professional development in formative assessment: experience in the USA

From 2003 to 2006 Wiliam, working at ETS in the USA with Leahy, a retired head teacher from England, experimented with ways of achieving the same effects as found in the KMOFAP project at a scale needed to reach large numbers of classrooms. They recognised that

> any model of effective, scalable teacher professional development would need to pull off a delicate balancing act between two conflicting requirements. The first was the need to ensure that the model was sufficiently flexible to allow the model to be adapted to the local circumstances of the intervention, not just to allow it to succeed, but also so that it could capitalize upon any affordances present in the local context that would enhance the intervention. The second was to ensure that the model was sufficiently rigid to ensure that any modifications that did take place preserved sufficient fidelity to the original design to provide a reasonable assurance that the intervention would not undergo a 'lethal mutation'.[114]

The researchers developed and piloted a number of models for working with teachers from which they drew conclusions about pace, length, duration and content of their interventions. For instance, they found that intervals of two weeks between meetings gave too little time for teachers to plan and implement changes and have something to report back at the next meeting. Monthly meetings were found to be optimum. Meetings of two hours duration were found to be too long in some cases but one hour meetings too short, so something in between was needed. In terms of the number of participants, between eight and 12 was found best, with group including teachers of a range of subject specialisms. It was also found that progress was helped by adopting a structure in the agenda for each meeting so that teachers knew what to expect and their role in the proceedings.

[111] Black, P et al (2003) op cit p 57

[112] Ibid, p 98.

[113] Wiliam, D. et al (2004) op cit

[114] Leahy, S. and Wiliam, D. (2012) op cit, p 54-5

The ideas arising from these trials were used to develop products for distribution to schools to support teachers in developing their formative assessment practice. Leahy and Wiliam developed two packs of materials, including video clips and agenda, handouts and notes for the group leader.[115] However, the researchers recognise that existence of these materials for professional development does not ensure their use, which depends on schools being willing to prioritize the implementation of formative assessment.

Transforming teachers' assessment practices in Scotland

In Scotland the formative assessment project was part of a bigger Assessment is for Learning (AifL) programme which considered the whole assessment system – pupils' records, personal planning, system monitoring and school evaluation, as well as formative assessment and summative assessment at the classroom level. The formative assessment project used a range of professional development activities including local workshops and national conferences at which presentations were made by teachers already involved in implementing formative assessment in England and later by Scottish teachers as the project progressed. There were also less formal, but nevertheless planned, discussions among local teachers and with Education Authority development officers. Some reference was also made to the publications of the Assessment Reform Group.[116] Although visiting researchers were included in the professional development programme there is a clear impression that their participation was decided, designed and controlled in Scotland, rather than being a formal collaboration with those developing formative assessment in England or elsewhere.

Whether or not the range of activities provided by the project was designed with the varying needs of different teachers in mind, it turned out to cater well for the several ways in which teachers came to implement formative assessment in their own practice. Three main approaches were identified

- *Trial and adjustment* : where teachers started by trying out strategies suggested in the publications or by teachers who were already using formative assessment. Strategies were adjusted in the light of experience during trials.
- *Critical review and discussion before trial :* ideas for implementing were discussed with colleagues to increase understanding before trial.
- *Starting from aims and ideas* : teachers developed ways of incorporating into their practice the ideas of formative assessment from their shared experience, rather as the initial group involved in the KMOFAP had done.

Although there was a general impression of successful implementation by those in the project, the report[117] comments on the apparent reluctance of some teachers to engage with theories of learning in order to understand why the strategies worked to enhance learning. It was also noted that very few teachers developed their practice to the extent of enabling students to take part in decisions about their learning goals. Although students were taking more initiative in solving problems and were reported as doing more thinking and being clearer about what they should be learning, the teachers maintained a grip on lesson objectives and targets. It may be that these two issues were related and that teachers enable students to take more responsibility for their learning when they themselves develop understanding of why formative assessment works in terms of how students learn. As the KMOFAP project showed, this takes time.

[115] Leahy, S. and Wiliam, D. (2009) *Embedding Assessment for Learning – A Professional Development Pack*. London: Specialist Schools and Academies Trust.
Leahy, S. and Wiliam, D. (2010) *Embedding Assessment for Learning – Pack 2*. London: Specialist Schools and Academies Trust.

[116] See www.assessment-reform-group.org/publications

[117] Hayward, L. & Spencer, E. (2010) The Complexities of Change: formative assessment in Scotland, *The Curriculum Journal*, 21 (2) 161-177. Routledge

The importance of inquiry in professional development in formative assessment

Looking across a number of research and development initiatives designed to help teacher develop formative assessment strategies in various countries, Pedder and James[118] identified classroom-based collaborative professional learning as the factor most strongly related to the use of some formative assessment strategies. Such collaborative learning can take different forms, from informal discussions among teachers, to teachers visiting each other's classrooms and discussing their observations, to the more formal Lesson Study developed in Japan. They also concluded that more emphasis needs to be placed on teachers having opportunities to use relevant research findings and to conduct research into their practice:

> If teachers are prepared and committed to engage in the risky business of problematising their own practice, seeking evidence to evaluate in order to judge where change is needed, and then to act on their decisions, they are thus engaging in assessment for learning (formative assessment) with respect to their own professional learning.[119]

Developing summative assessment by secondary teachers

In an attempt to find an approach to summative assessment that would have fewer negative effects on teachers and students than the national tests in operation in England, researchers and advisers from two local authorities worked with teachers to develop the use of teachers' judgements for summative assessment[120]. The aim was to find methods and processes for ensuring the comparability of judgements between teachers and schools. In its pilot phase, the project worked with a small group of English and mathematics teachers of Year 8 students (aged 13). The teachers selected were all well versed in formative assessment practices. The project focused on the process of making summative judgements and how evidence is turned into judgements. The development was initiated by teachers identifying 'what does it mean to be good at this subject for Year 8 students?' In the second phase during the school year 2005-2006 new practices were tried out and adapted. The third phase involved the teachers spreading the ideas and practices across their departments.

The findings of the pilot project, derived from field notes, class observations, interviews and records of meetings indicated some confusion between formative and summative assessment and acceptance rather than challenge of the quality of current tests. There were also differences in the reactions of teachers of mathematics and English. Teachers of English were reported as being comfortable with a portfolio system, placing greater emphasis on speaking and listening and introducing a 'controlled' piece of work, on which students worked alone. Mathematics teachers preferred to keep to the use of tests (some queried the need for change) but to improve them, or to introduce 'alternative assessment tasks' rather than a more holistic approach that was favoured by the English teachers. However, the mathematics approach was found not to provide pupils opportunities to show a range of achievement, which came as a surprise to the teachers when they became aware of it.

Developing summative assessment by primary teachers

Since 1995 teachers in England have been expected to assess their students in the final year (Year 6, age 11) of primary school as well as administering national tests in English, mathematics and science. In theory the assessment by teachers was intended to provide data about a wider range of aspects of the subjects than could be included in the tests. In practice, many teachers based their teachers' assessment on giving

[118] Pedder, D. and James, M. (2012) Professional learning as a condition for assessment for learning. In (ed) J. Gardner *Assessment and Learning*. 2nd edn. London: Sage, pp 33-48.

[119] ibid p 41/2

[120] Black, P., Harrison, C., Hodgen, J., Marshall, B. and Serret, N (2006) Riding the interface: an exploration of the issues that beset teachers as they strive for assessment systems. BERA conference paper see http://www.kcl.ac.uk/sspp/departments/education/research/crestem/assessment/riding.pdf

their students past national tests, thus defeating the purpose of reporting their judgements and losing the opportunity to use data collected in regular work formatively as well as summatively.

In 2006 a small project[121] was begun in an attempt to find an alternative model for statutory end-of-primary school teachers' assessment and moderation. To this end meetings were held with a small group of Year 6 teachers during the time when they were in the process of conducting their end-of-year judgements. These judgements had to be reported by deciding the 'level' achieved by each student. 'Levels' are identified in the National Curriculum Attainment Targets for each subject. (See the example in Box 19, p 64.) The meetings enabled teachers to reflect on their experience of different ways of arriving at decisions about levels and to moderate the processes. It became clear in the first meetings that teachers were using a narrow evidence base for their judgements, possibly a result of emulating tests in deciding what to take into account. Making judgements of levels was seen as problematic since the level descriptions span two years of work. There was a desire to create sub-levels but recognition that it was difficult to establish firm evidence for intermediate steps between the levels. Instead, working with the advisers, teachers produced some proposals for expanding the range of evidence used and some principles for guiding practice in making judgements about levels achieved.

These materials were made available to a larger group of year 6 teachers from 24 schools who took part in the project in 2007. The teachers collected evidence over a period of two terms and brought examples to two meetings where their judgements were moderated in discussion with other teachers. The meetings not only gave them direct experience of moderation but opportunity to share their experiences and then discuss their selections of evidence. They widened the evidence base of their judgements and began to see that pupils might have a role in their own assessment. The reaction of teachers to the experience of collaborative moderation was highly positive.[122] They thought that all teachers should have such opportunities. Whilst those involved valued experience for themselves of developing guidance material they suggested that, for wider dissemination the process would have to be 'formalised', which would inevitably make it more 'top down' for others.

Changing pedagogy and assessment practices in France

The considerable changes in French education introduced between 2006 and 2009, mentioned in Chapter 6 (page 67), required schools to implement a common core of skills as well as knowledge. The specification of skills was a novelty and in science required some teachers to change their approach to teaching at the same time as their assessment practices. The *La main à la pâte* project has provided help for schools with both of these changes appropriately to the existing practices of the schools. In one district primary school teachers were helped to develop teaching modules on a particular theme making progressive use of inquiry skills for pre-school, grades 1 and 2, and grades 3 to 5.[123] Each module contains a section on assessment. Tests were also constructed and given to a sample of students at the end of primary school (grade 5) in order to assess their skills. The performance was reported not as marks or scores but as 'working through inquiry', 'progressing towards inquiry' and 'not yet working through inquiry'. Unpublished reports show that over three years the number of students not working through inquiry decreased from 29% to 8%.

In another district, where IBSE was already being implemented, schools were given access to tasks such as the one described in Example 7 (page 68) to enable them to assess their students' progress in inquiry skills and knowledge. Although the tests were not designed as summative assessment, teachers used the results as well as their observations of students' actions and other evidence in order to complete the record booklet for each student at the end of the year. The effect of providing access to

[121] Birmingham City Council Advisory and Support Service (2005) *Effective Assessment*

[122] Harlen, W. (2010) Professional learning to support teacher assessment, in Gardner, J. et al *Developing Teacher Assessment*. Maidenhead, England: Open University Press page 100-129.

[123] (see http://www.crdp-montpellier.fr/cd66/map66/projets_federatifs/air/index.php).

the tasks and to the guide for using evidence from students' notebooks, presentations and actions (see page 67) was slow to appear. Not much change was observed in the first year but thereafter the use of the assessment tasks and guidance spread rapidly to a large proportion (an estimated 75%) of the schools in the district.

Developing assessment practices through curriculum materials in Australia[124]

In Australia, the *Primary Connections* project aims at developing students' knowledge, skills, understanding and capacities in both science and literacy. It promotes an inquiry approach through a version of a five phase lesson structure, 5Es: Engage, Explore, Explain, Elaborate and Evaluate. The development of the project was funded by the Australian Government from 2004-2012, and it is used in one way or another by 56% of Australian primary schools. The programme has two main components: a professional learning programme and a suite of thirty-one curriculum units which cover the Australian science curriculum from Foundation year to Year 6. The curriculum units model inquiry practices and assessment for and of learning. The professional learning programme includes a workshop on assessment, and several training DVDs have been produced to provide examples of practice in classrooms.

Strategies are suggested which can provide information for both formative and summative assessment. Sets of teacher questions are included in lesson steps to elicit students' thinking and make their ideas accessible to teachers and students so learning can be monitored. Strategies for peer- and self-assessment are also included. The appendix for each unit summarises the assessment opportunities for each phase and assessment rubrics are available on the website to assist teachers to monitor students' conceptual understanding and inquiry skills.

A report of feedback from trial teachers[125] notes that the purposes of the Evaluate phase are to:

- provide an opportunity for students to review and reflect on their learning and new understanding and skills
- provide evidence for changes to students' understanding, beliefs and skills.

In relation to the first there was strong evidence that students reviewed their conceptual understanding but not their skills, although other evidence showed that they used a variety of skills. Teachers' responses in relation to the second purpose were sparse and suggested that this aspect of the phase may not have been implemented by many teachers. Those who did, employed a range of strategies which appeared to serve both purposes of the Evaluate phase, but again only in relation to conceptual understanding. The strategies included quizzes, writing a newspaper article, creating an imaginary animal based on criteria, drawings, diagrams, word loops, concept maps, concept cartoons, role plays and presentations. Teachers noted that a range of tasks, including individual questioning and novel situations, was necessary to assess students' understanding, all of which make heavy demands on teachers' time.

Overall, while teacher assessment has not been strong, the research evidence shows that some teachers are gradually moving towards better assessment practices and are building their confidence. The procedures found effective in improving teachers' assessment practices include:

- curriculum resources that embed research-based assessment examples
- professional learning which provides the theory behind why the approach works
- instructional DVDs which show what it looks like in real classrooms
- positive teaching experiences where student enjoyment and evidence of learning leads to teacher enjoyment and motivation to engage with change in classroom practice.

[124] Primary Connections http://science.org.au/primaryconnections/

[125] Skamp, K. (2012) Trial-teacher feedback on the implementation of Primary Connections and the 5E model. Australian Academy of Science http://www.science.org.au/primaryconnections/research-and-evaluation/teaching-ps.html

Implications for the assessment of IBSE

There are some lessons to be learned from these attempts to change assessment that can apply to learning and teaching in any subject. Before considering what may be needed in the case of assessing inquiry-based science education it is useful to recall that, since not all science teaching will involve inquiry, there should be a variety of assessment tasks and procedures to match the variety of different learning goals. As mentioned before, there is a place for direct instruction in conventions, vocabulary and basic knowledge which can be checked by teacher-made tests or quizzes. But all assessment must not be of the kind where answers can be recalled. Otherwise there will be no information about the extent of students' inquiry skills and scientific understanding. Our focus here is on those essential elements of assessment that deal with the goals of IBSE.

Unfortunately the use of assessment procedures that are fitted to the purposes of IBSE goals is not widespread and there is hardly any experience of what is needed to change assessment practice to fit better these purposes. To provide data about the range of competences that are the goals of IBSE requires a corresponding range of assessment procedures - some involving observing students' actions, others enabling students to express their ideas in discussion as well as in writing and all providing the opportunity for students to use scientific inquiry skills and reasoning. As noted in Chapter 6, examples of such procedures exist but are not widely used in science education. However, by extrapolation from procedures found successful in other domains, it is possible to find some pointers to what helps teachers to change their assessment practice in science.

There seem to be four key elements in bringing about change in assessment practices at class and school levels which apply both to formative and summative assessment:

- Motivation to change
- Goals to aim for
- Opportunities for teachers to discuss, compare and share solutions
- Means to evaluate changed practice.

Motivation to make change comes from discontent with current practice and recognising that there can be improvement. In the experience of the KMOFAP project (page 74) a key factor in persuading teachers to make the effort to introduce formative assessment was research evidence of the impact on students' achievement. A factor in keeping it going once started was the evidence of impact of changes in their teaching on their own students. In the case of summative assessment, motivation to change can come from providing teachers with evidence that conventional summative assessment by conventional testing is not giving their students opportunities to show what they can do and understand. Engaging teachers with more research of the kind undertaken by Dolin and Krogh, noted in Chapter 6 (page 60), could be influential. Evidence that the assessment procedure can make a significant difference is a first step to teachers identifying that it is the view of learning behind assessment methods that makes the difference. An approach that brings assessment more into line with the socio-cultural setting that supports IBSE enables students to show their capabilities to a far greater extent than individual tests or tasks.

Goals to aim for means that teachers and schools have a clear idea of what kinds of change will be needed but have to develop for themselves the detailed steps needed to bring about that change suited to their particular circumstances. In the case of formative assessment teachers may be given information about the range of effective strategies but adopt different ways implementing them in their particular classroom. In changing summative assessment, teachers can be helped to refine their ideas of how achievement and progress are identified and how to ensure that students have the opportunities to show their achievement and progress.

Opportunities to discuss, compare, and share solutions ensure that, although tailored to particular situations, practices maintain a focus on common goals. Discussion with others attempting to achieve the same effects but in different contexts provides teachers with ideas which may not have occurred to them working alone. Collaborative inquiry operates to help teachers to construct and reconstruct ideas just as it does for students. It is important that this collaboration and shared thinking is available to all. Some of the projects that began with teachers' collaborative development of new assessment practices led to the production of materials that teachers thought to be needed by others. In such cases, what begins as transformation for the few ends as transmission to the many. Avoiding this requires the 'delicate balancing act' identified by Leahy and Wiliam, to avoid the pitfall of transmission models of change – the assumption that practice can be changed by following a recipe provided by others, even if these others are teachers.

Means to evaluate changes in assessment practice enables teachers to continue to develop and to adjust their practice to new circumstances. What ought not to change (at least only in the long term) are the principles which guide and provide the justification for decisions. Unless reasons for change are understood techniques have to be followed somewhat blindly; they will not be adapted as necessary for different situations and will eventually become less useful. It may well be, as in the case of the KMOFAP project teachers (page 74), that interest in knowing *why* new practices work develops after seeing that they *do* work. However, whether the underlying rationale is presented sooner or later in the process of making changes, at some point teachers need to have a reason for adopting new and abandoning old practices.

As an example of principles relating to assessment practice, the list in Box 21 emerged from a project which studied the processes involved in making changes in assessment practice, particularly in placing the role of teachers firmly at the centre of assessment.

The last of these principles concerns standards of quality that can be used to evaluate practice. The project which identified the principles also proposed some standards to be met by those making decisions about assessment in classrooms, at school level, at local authority level and as part of national educational policy. For the individual student, what happens in the classroom has the greatest impact on learning; thus there are standards to be met if the assessment is to help learning. But what happens in the classroom is dependent upon what happens at the school level in terms of assessment policy and general discourse about how assessment ought to help or to report or to measure students' achievements. In turn school policy and practice is influenced by local authority guidance and by national policy and requirements. Hence the project identified standards to be met by practice within each of these communities. The standards are reproduced in Tables 1 - 4. In each case there are standards for assessment generally, for formative assessment and for summative assessment.

[126] Gardner, J., Harlen, W., Hayward, L. and Stobart, G. with Montgomery, M. (2010). *Developing Teacher Assessment.* Maidenhead: Open University Press pp 48-51.

Box 21: Principles of assessment practice

1 Assessment of any kind should ultimately improve learning.

2 Assessment methods should enable progress in all important learning goals to be facilitated and reported.

3 Assessment procedures should include explicit processes to ensure that information is valid and is as reliable as necessary for its purpose.

4 Assessment should promote public understanding of learning goals relevant to students' current and future lives.

5 Assessment of learning outcomes should be treated as approximations, subject to unavoidable errors.

6 Assessment should be part of a process of teaching that enables students to understand the aims of their learning and how the quality of their achievement will be judged.

7 Assessment methods should promote the active engagement of students in their learning and its assessment.

8 Assessment should enable and motivate students to show what they can do.

9 Assessment should combine information of different kinds, including students' self-assessments, to inform decisions about students' learning and achievements.

10 Assessment methods should meet standards that reflect a broad consensus on quality at all levels from classroom practice to national policy.[126]

Table 1: Standards for Classroom Assessment Practice

Assessment Generally	Formative Use of Assessment	Summative Use of Assessment
1 The assessment uses a range of methods that enable the various goals of learning and progression towards them to be addressed 2 The methods used address the skills, knowledge or understanding being assessed without restricting the breadth of the curriculum 3 Teaching provides students with opportunities to show what they can do through tasks that address the full range of goals of learning 4 Teachers use evidence from their on-going assessment to: • help students' learning; • summarize learning in terms of reporting criteria; • reflect upon and improve their teaching 5 Teachers develop their assessment practice through a variety of professional learning activities including reflecting upon and sharing experiences with colleagues	1 Teachers gather evidence of their students' learning through questioning, observation, discussion and study of products relevant to the learning goals 2 Teachers involve students in discussing learning goals and the standards to be expected in their work 3 Teachers use assessment to advance students' learning by: • adapting the pace, challenge and content of activities • giving feedback to students about how to improve • providing time for students to reflect on and assess their own work 4 Students use assessment to advance their learning by: • knowing and using the criteria for the standards of work they should be aiming for • giving and receiving comments from their peers on the quality of their work and how to improve it • reflecting on how to improve their work and taking responsibility for it	1 Teachers base their judgements of students' learning outcomes on a range of types of activity suited to the subject matter and age of students, which might include tests or specific assessment tasks 2 Assessment of learning outcomes is based on a rich variety of tasks that enables students to show what it means to be 'good' at particular work 3 Teachers take part in discussion with each other of students' work in order to align judgements of levels or grades when these are required 4 Students are aware of the criteria by which their work over a period of time is judged 5 Students are aware of the evidence used and how judgements of their learning outcomes are made 6 Students are helped to use the results of assessment to improve their learning

Table 2: Standards for Use by School Management Teams

Assessment Generally	Formative Use of Assessment	Summative Use of Assessment
1 There is a school policy for assessment that reflects the standards above for classroom practice 2 The policy is regularly discussed and reviewed to reflect developing practice 3 Teachers have opportunities to improve their assessment practice through professional learning and collaboration 4 Time is made available for teachers to discuss, reflect upon and on occasion to observe each others' assessment practice 5 The school's policy and practice in assessment are communicated to parents and carers	Teachers collaborate in developing their practice in: • communicating goals and quality criteria to students • helping students to take part in self- and peer-assessment • providing feedback to help learning • enabling students to take responsibility for their work	1 Teachers are able to use a variety of assessment methods free from the pressure of high stakes use of the results 2 Teachers take part in developing quality assurance procedures to maximize consistency in their judgements 3 Students' achievements are discussed in terms of what they can do and not only in terms of levels or grades 4 A manageable system for record-keeping is in operation to track and report on students' learning 5 Parents and carers receive written and oral reports that identify the next steps for their children and provide information about assessment processes to ensure confidence in teachers' assessment 6 Summative judgements are required only when necessary to check and report progress

Table 3: Standards for Use in National and Local Inspection and Advice Arrangements

Assessment Generally	Formative Use of Assessment	Summative Use of Assessment
1 Schools' policies and practices in assessment are reviewed in relation to the standards above 2 Inspection procedures ensure that schools evaluate their assessment practices and develop action plans for improvement 3 There are opportunities for schools to share and develop assessment practices 4 Professional development is available to develop policies and improve assessment practice 5 Resources are available to enable schools to take steps to improve assessment practice	1 Schools' use of assessment to support learning is included as a key factor in evaluating the effectiveness of schools 2 Help is available for schools to ensure that all areas of achievement benefit from the formative use of assessment 3 Schools are encouraged to develop their formative use of assessment	1 Schools are helped to develop action plans based on self-evaluation across a range of indicators beyond students' levels of achievement 2 Advice on school assessment policies and practices takes account of what is known about the reliability and validity of different assessment methods 3 Schools are helped to use assessment results to identify areas for improvement of learning opportunities

Table 4: Standards for Use in National Policy Formulation

Assessment Generally	Formative Use of Assessment	Summative Use of Assessment
1 Policies require schools and local advisers to show how all assessment is being used to help students' learning	1 Assessment to support learning is at the heart of government programmes for raising standards of achievement	1 Moderated assessment by teachers is used to report students' performance throughout the compulsory years of school
2 Introduction of new practices in assessment is accompanied by changes in teacher education and evaluation criteria necessary for their sustainability	2 Initial teacher education and professional development courses ensure that teachers have the skills to use assessment to support learning	2 Moderation of teachers' judgements is required to ensure common interpretation of criteria within and across schools.
3 Schools are accountable for using formative and summative assessment to maximize the achievement of goals	3 School inspection frameworks give prominence to the use of assessment to support learning	3 Regulations ensure that arrangements for the summative use of assessment are compatible with the practice of using assessment to help learning
4 National standards of students' achievement are reported as a range of qualitative and quantitative data from surveys of representative samples	4 Schools are encouraged to evaluate and develop their formative use of assessment	4 Targets for school improvement are based on a range of indicators and are agreed through a process combining external evaluation and internal self-evaluation

Bibliography

Alexander, R. (Ed) (2010) *Children, their World, their Education*. Final report and recommendations of the Cambridge Primary Review. London: Routledge.

Alexander, R. (2004) *Towards Dialogic Teaching. Rethinking Classroom Talk*. Cambridge: Dialogos.

Assessment Reform Group (ARG) (2002) *Assessment for Learning: 10 Principles*. www.assessment-reform-group.org

Australian Government Department Education, Employment and Workplace Relations. *Science Education Assessment Resources (SEAR)* http://cms.curriculum.edu.au/SEAR

Barnes, D. (1976) *From Communication to Curriculum*. Harmondsworth: Penguin.

Birmingham City Council Advisory and Support Service (2005) *Effective Assessment.*

Black, P. (1998) *Testing: Friend or Foe?* London: Falmer Press.

Black, P., Harrison, C., Hodgen, J., Marshall, B. and Serret, N (2006) Riding the interface: an exploration of the issues that beset teachers as they strive for assessment systems. BERA conference paper . See http://www.kcl.ac.uk/sspp/departments/education/research/crestem/assessment/riding.pdf

Black, P., Harrison, C., Lee, C., Marshall, B. and Wiliam, D. (2003). *Assessment for Learning: Putting it into Practice.* Maidenhead England: Open University Press.

Black, P. and Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21 (1). 5-13.

Black, P. and Wiliam, D. (1998) Assessment and classroom learning, *Assessment in Education,* 5 (1) 7-74.

Bransford, J.D., Brown, A. and Cocking, R.R. (eds) (2000) *How People Learn, Brain, Mind, Experience and School.* Washington, D.C.: National Academy Press.

Bruner, J. (1996) *The Culture of Education.* Cambridge, MA: Harvard University Press.

Budd-Rowe, M. (1974) Relation of wait-time and rewards to the development of language, logic and fate control: Part II, *Journal of Research in Science Teaching,* 11(4) 291-308.

Butler, R. (1988) Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest and performance, *British Journal of Educational Psychology* 58, 1-14.

Crossouard (2012) Absent presences: the recognition of social class and gender dimensions within peer assessment interactions, *British Educational Research Journal*, 38 (5) 731-748.

Department for Education (2011). *The Framework for the National Curriculum. A report by the Expert Panel for the National Curriculum review*. London: Department for Education.

DES, DENI and WO (1985) *APU Science in Schools Age 11 Report no 4*. London: HMSO.

DES and WO (1988) *National Curriculum Task Group on Assessment and Testing: A Report*. London: HMSO.

DES, DENI and WO (1981) *Science in Schools Age 11 Report no 1*. London: HMSO.

Dewey, J. (1933) *How we think: A restatement of the relation of reflective thinking to the educative process*. Boston, MA: D.C. Heath.

Dolin, J., & Krogh, L. B. (2010): The Relevance and Consequences of Pisa Science in a Danish Context. *International Journal of Science and Mathematics Education,* 8, 565-592.

Dolin, J., Laursen, E., Raae, P. H., Senger, U. (2005). *Udviklingsprojekter som læringsrum. Potentialer og barrierer for skoleudvikling i det almene gymnasium. Gymnasiepædagogik nr. 54,* Syddansk Universitet. 232 s. (Development projects as learning arenas. Potentials and barriers for school development in the general upper secondary school. University of Southern Denmark). http://www.sdu.dk/Om_SDU/Institutter_centre/Ikv/Formidling/Tidsskrifter/Gymnasiepeadagogik/Udgaver Accessed Dec11, 2012.

Ertl, H. (2006) Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education,* 32 (5) pp 619–634.

Gardner, J., Harlen, W., Hayward, L. and Stobart, G. with Montgomery, M. (2010). *Developing Teacher Assessment*. Maidenhead: Open University Press.

Gipps, C., McCallum, B. and Brown, M. (1996). Models of teacher assessment among primary school teachers in England, *The Curriculum Journal,* 7 (2) 167-183.

Grigorenko, E. (1998) Mastering tools of the mind in school (trying out Vygotsky's ideas in classrooms), in (eds) R. Sternberg and W. Wiliams *Intelligence, Instruction and Assessment: Theory and Practice*. Mahwah, NJ: Erlbaum.

Harlen, W. (2012a) The role of assessment in developing motivation for learning, in (ed) J. Gardner *Assessment and Learning*. London: Sage pp171-184.

Harlen, W. (2012b) On the relationship between assessment for formative and summative purposes, in (ed) J. Gardner *Assessment and Learning*. London: Sage pp 87-102.

Harlen, W. (2010) Professional learning to support teacher assessment, in Gardner, J. *et al Developing Teacher Assessment*. Maidenhead England: Open University Press page 100-129.

Harlen (Ed) (2010) *Principles and Big Ideas of Science Education*. Available from www.ase.org.uk in English, from www.fondation-lamap.org in French, and from www.innovec.org.mx in Spanish.

Harlen, W. (2007) *Assessment of Learning*. London: Sage.

Harlen, W. (2006) *Teaching, Learning and Assessing Science 5 – 12*. 4th edn. London: Sage.

Harlen, W. (2004) Trusting teachers' judgements: research evidence of the reliability and validity of teachers' assessment for summative purposes, *Research Papers in Education*, 20(3); 245-270.

Harlen, W. and Qualter, A. (2009) *The teaching of Science in Primary Schools*. London: Routledge.

Hattie, J. and Timperley, H. (2007) The power of feedback. *Review of Educational Research*, 77, 81-112.

Hayward, L. (2010) Moving beyond the classroom, in J. Gardner et al *Developing Teacher Assessment*. Maidenhead, UK: Open University Press.

Hayward, L. & Spencer, E. (2010) The Complexities of Change: formative assessment in Scotland, *The Curriculum Journal,* 21 (2) 161-177.

IAP (2012) Taking Inquiry-Based Science Education into Secondary Education. Report of a global conference. http://www.sazu.si/files/file-147.pdf

Jager, J.J., Merki, K.M., Oerke, B. and Holmeier, M. (2012) State-wide low-stakes tests and a teaching to the test effect? An analysis of teacher survey data from two German States, *Assessment in Education,* 19 (4) 451-467.

James, M. (2012) Assessment in harmony with our understanding of learning: problems and possibilities, in (ed) J. Gardner *Assessment and Learning*, 2nd edn. London: Sage 187 – 205.

Leahy, S. and Wiliam D. (2012) From teachers to schools: scaling up professional development for formative assessment, in (ed) J. Gardner *Assessment and Learning*, 2nd edn. London: Sage. 49-71.

Leahy, S. and Wiliam, D. (2009) *Embedding Assessment for Learning – A Professional Development Pack.* London: Specialist Schools and Academies Trust.

Leahy, S. and Wiliam, D. (2010) *Embedding Assessment for Learning – Pack 2.* London: Specialist Schools and Academies Trust.

Linn, R. L. (2000) Assessments and accountability, *Educational Researcher,* 29 (2) 4-16.

Masters, G. and Forster, M. (1996) *Progress Maps.* Camberwell, Victoria, Australia: ACER.

Maxwell, G. (2004) 'Progressive assessment for learning and certification: some lessons from school-based assessment in Queensland.' Paper presents at the third conference of the Association of Commonwealth Examination and Assessment Boards, March Nidi, Fiji.

Messick, S.(1989) Validity, in (ed) R. Linn *Educational Measurement* (3rd Edn)American Council on Education , Washington: Macmillan, pp 13-103.

Michaels, S., Shouse, A.W. and Schweingruber, H.A (2008) *Ready, Set, Science! Putting research to work in K-8 Science Classrooms,* Washington: National Academies Press.

Minner, D.D., Levy, A. J and , Century, J. (2010) Inquiry-Based Science Instruction—What Is It and Does It Matter? Results from a Research Synthesis Years 1984 to 2002*, Journal of Research in Science Teaching*, 47 (4) 474-496.

National Research Council (2012) *A Framework for K-12 Science Education.* Washington DC: National Academies Press.

Newton, P. E. (2012) Validity, purpose and the recycling of results from educational assessment, in (Ed) J. Gardner *Assessment and Learning* 2nd edition. London: Sage 264-276.

Noble, T., Suarez, C., Rosebery, A., O'Connor, M.C. ,Warren, B. and Hudicourt-Barnes, J. (2012) ''I never thought of it as freezing'': How Students Answer Questions on large-scale science tests and what they know about science, *Journal of Research in Science Teaching*, 49 (6) 778–803.

Nordenbo, S. E., Allerup, P., Andersen, H. L., Dolin, J., Korp, H., Larsen, M. S., et al. (2009). *Pædagogisk brug af test - Et systematisk review*. København: Aarhus Universitetsforlag. (In English: *Pedagogical use of tests – A systematic review*).

Nuffield Primary Science Teachers' Guide *Materials*. (1995) London: Collins Educational.

Nusche, D., Laveault, D., MacBeath, J. and Santiago, P. (2012) *OECD Reviews of Evaluation and Assessment in Education: New Zealand 2011.* Paris: OECD.

OECD (2011) *Towards an OECD Skills Strategy.* Paris: OECD.

OECD (2006) *PISA released items: Science*. Paris: OECD http://www.oecd.org/pisa/38709385.pdf

OECD 2003, *The PISA 2003 Assessment Framework* Paris: OECD.

OECD (2000) *Measuring Student Knowledge and Skills: A new Framework for Assessment*. Paris: OECD.

Osmundson, E., Chung, G., Herl, H., Klein D. (1999) *Knowledge-mapping in the classroom: a tool for examining the development of students' conceptual understandings*. Los Angeles, California: National Centre for Research on Evaluation and Student Testing, University of California. www.cse.ucla.edu/Reports/TECH507.pdf

Osborne, J., Simon, S. and Collins, S.(2003) Attitudes towards science: a review of the literature and its implications*, International Journal of Science Education*, 25, 1049-1079.

Pedder, D. and James, M. (2012) Professional learning as a condition for assessment for learning. In (ed) J. Gardner *Assessment and Learning.* 2nd edn. London: Sage, pp 33-48.

Pellegrino, J.W., Chudowsky, N. and Glaser, R. (Eds) (2001) *Knowing what Students Know The Science and Design and Educational Assessment*. Washington, DC: National Academy Press.

Piaget, J (1929) *The Child's Conception of the World*. New York: Harcourt Brace.

Pine, J., Aschbacher, P., Rother, E., Jones, M., McPhee. C., Martin, C., Phelps, S., Kyle, T. and Foley, B. (2006) Fifth graders' science inquiry abilities: a comparative study of students in hands-on and textbook curricula, *Journal of Research in Science Teaching* 43 (5): 467-484.

Pollard, A and Triggs, P. (2000) *Policy, Practice and Pupil Experience*. London: Continuum.

Pollard, A., Triggs, P., Broadfoot, P., Mcness, E. and Osborn, M. (2000) *What pupils say: changing policy and practice in primary education*. London: Continuum.

Primary Connections http://science.org.au/primaryconnections/

Pryor, J. and Lubisi, C. (2001) Reconceptualising educational assessment in South Africa –testing times for teachers, *International Journal for Educational Development,* 22 (6), 673-686.

Roderick, M. and Engel, M. (2001) The grasshopper and the ant: motivational responses of low achieving pupils to high stakes testing. *Educational Evaluation and Policy Analysis* 23: 197-228.

Sadler, D. R. (1989) Formative assessment and the design of instructional systems, *Instructional Science*, 18, 119-44.

SEED (Scottish Executive Education Department) (2002) *How Good is Our School? Self evaluation using quality indicators*. Edinburgh: HMIE.

Skamp, K. (2012) *Trial-teacher feedback on the implementation of Primary Connections and the 5E model*. Australian Academy of Science. http://www.science.org.au/primaryconnections/research-and-evaluation/teaching-ps.html

Stobart, G. (2012)Validity in formative assessment, in (ed) J. Gardner *Assessment and Learning*. 2nd edn. London: Sage pp 233-242.

Stobart, G. (2008) *Testing Times. The uses and abuses of assessment*. London: Routledge.

Streeter, L., Bernstein, J., Foltz, P. and DeLand, D. (2011) *Pearson's Automated Scoring of Writing, Speaking, and Mathematics*: Pearson http://kt.pearsonassessments.com/download/PearsonAutomatedScoring-WritingSpeakingMath-051911.pdf

Tymms, P. (2004) Are standards rising in English primary schools? *British Educational Research Journal*, 30 (4) 477-94.

Vygotsky, L.S. (1978) *Mind in Society: The Development of Higher Psychological Process*. Cambridge, MA: Harvard University Process.

Watkins, C. (2003) *Learning: A Sense-Maker's Guide*. London: Association of Teachers and Lecturers.

Welford, G., Harlen, W. and Schofield, B. (1985) *Practical Testing at ages 11, 13, and 15*. London: DES, WO and DENI.

White, R. T. (1988) *Learning Science*. Oxford: Blackwell.

Wiliam, D. (2001) Reliability, validity and all that jazz, *Education 3-13*, 29 (3): 17-21.

Wiliam, D. (2009) An integrative summary of the research literature and implications for a new theory of formative assessment, in (eds) H. L. Andrade and G. J. Cizek, *Handbook of Formative Assessment*, New York: Taylor and Francis.

Wiliam, D. Lee, C., Harrison, C. and Black, P. (2004) Teachers developing assessment for learning: impact on student achievement, *Assessment in Education*, 11 (1) 49-66.