# A review of current thinking and practices in assessment in relation to the Primary Years Programme

Wynne Harlen
Sandra Johnson

Report submitted by Assessment Europe to the International Baccalaureate
January 2014

# Contents

# Executive summary

The principal purpose of the research project described in this report was to contribute to a substantial review of the Primary Years Programme (PYP), by critically reflecting on the PYP curriculum framework in the light of the approaches, strategies and tools of assessment of 21st century primary education. The PYP curriculum offers an inquiry-based trans-disciplinary learning experience to students aged 3 to 12 years, with a focus on the development of concepts, knowledge, skills, attitudes and action. The programme culminates with the 'PYP exhibition', a collaborative venture in which students work together on a topic chosen in consultation with their teacher, and eventually present the results of their efforts to fellow classmates, teachers and the wider school community

The assessment review focused both on ongoing formative assessment within the PYP (assessment for learning) and summative assessment of individual and group performances in the exhibition (assessment of learning). It was supported by findings from an extensive review of the literature, which covered writings on assessment published during the past five years, including theoretical pieces on the nature of assessment, i.e. the roles assessment can take in teaching and learning and how these are best practised to support learning of concepts, knowledge, skills, attitudes and action supporting the cognitive, affective and behavioural goals of primary education in the 21st century, as well as reports on assessment practice with students aged 3 to 12.

The key findings and implications for assessment in the PYP that emerged from the review are as follows:

- There is solid and substantial support in academic literature and research for the emphasis given to formative assessment in the PYP approach to student assessment.
- Formative assessment should be presented as integral to effective teaching rather than a process that is separate from, but can be linked to, teaching.
- The key features of formative and summative assessment should be clearly distinguished.
- Teachers should be encouraged to extend the range of formative assessment strategies they use beyond feedback to students.
- Attention should be given to teachers helping students to recognise the goals of their work and take part in decisions about what they need to do to achieve them.
- The PYP should consider providing a set of general criteria statements constituting a template to be used by teachers to identify criteria appropriate to particular exhibition and unit topics.
- Group moderation of teachers' judgements should be promoted, where teachers discuss their judgements of students' achievements in relation to the assessment criteria.

# 1. Introduction

*This chapter introduces, in section 1.1, the purpose of the review and the questions it was designed to address. As background information for those not familiar with the PYP, sections 1.2 and 1.3 provide a brief description of the PYP curriculum and the role of student assessment within it. Readers already familiar with these features of the programme might move directly to section 1.4.*

## 1.1 The purpose of this review

The principal purpose of the research project described in this report was to contribute to a substantial review of the Primary Years Programme (PYP), by reviewing the assessed PYP curriculum framework in the light of the approaches, strategies and tools of assessment of 21st century primary education.

The review addresses the following broad research questions:

### Current thinking and practices

- What are the current thinking and effective practices suggested by research around assessment in primary education in different educational contexts (students aged three to twelve)?

- What are key principles, approaches, models, strategies and tools of assessment that impact most profoundly on students' learning?

- How do these strategies and tools measure, evaluate, monitor and support student learning outcomes and progress continuously (including learning goals, readiness, needs, preferences and interests)?

### PYP reflecting current thinking and practice

- To what extent do these best practice assessment structures, objectives, strategies and tools currently reflect the overall philosophy and pedagogical approach of the PYP and feature within its curriculum framework?

### PYP Exhibition

- How could the PYP exhibition be planned and used in effective and creative ways to support assessment reflecting the achievement in the learning of the five elements embedded in the PYP framework (that is, concepts, knowledge, skills, attitudes and action)?

- How could these approaches be used by PYP schools to measure, evaluate, monitor, and support student learning outcomes and progress in relation to each of these five essential elements?

### Implications, opportunities and challenges for assessment in the PYP

- Based on the outcomes of the above questions, what are key implications, opportunities and challenges for the development of assessment in the programme?

The review comprised two major components:

1. A two-strand review of the academic literature reporting practice in assessment for, of and as learning in cognitive, affective and behavioural domains in different education systems, following different educational paradigms and approaches, with a particular emphasis on 21st century primary education:

   *Strand 1*
   The nature of assessment, the roles it can take in teaching and learning and how these are best practised to support learning of concepts, knowledge, skills, attitudes and action supporting the cognitive, affective and behavioural goals of primary education in the 21st century. The review was to set out the relative merits of major assessment procedures such as tests and examinations, teacher-based assessment, special tasks, and so on, and address the relationship between views of learning underpinning the pedagogy and content and the various assessment procedures.

   *Strand 2*
   A review of practices in assessment in primary and lower secondary education (pupils aged 3 to 12) in various educational systems. This was to include practice in using assessment to help learning (formative assessment, or assessment for learning), to report on learning (summative assessment, or assessment of learning) and to identify assessment as an integral part of learning (assessment as learning). On the basis of this review those practices considered to provide the most valid and reliable evidence for various purposes would be identified.

2. An evaluation of current assessment policy and practice within the PYP, based on salient findings from the literature review about best practice in different contexts, and bearing in mind transition to the MYP and eventually to the Diploma.

## 1.2 The PYP curriculum

The PYP curriculum framework offers a transdisciplinary learning experience to students aged 3 to 12 years, with a focus on the development of the 'essential elements': concepts, knowledge, skills, attitudes and action. Learning is principally through inquiry within a framework that focuses on six transdisciplinary themes (IB 2009): 'Who we are', 'Where we are in space and time', 'How we express ourselves', 'How the world works', 'How we organize ourselves' and 'Sharing the planet'. The youngest students, the 3 -5 year olds, complete four study units per year, which must include 'Who we are' and 'How we express ourselves'; from age 5 upwards students study six units covering all six themes. In the final year of the programme one unit is replaced by the 'PYP exhibition', a collaborative venture in which students work together on a topic chosen in consultation with their teacher, and eventually present the results of their efforts to fellow classmates, teachers and the wider school community (for example, the governing body, parents, secondary school colleagues and students).

It is this collaborative inquiry-based approach to learning, combined with the broad aim to develop students' skills in all three of the cognitive, behavioural and affective domains, and the aim to develop learners described by the learner profile, that characterises the PYP framework as delivering a '21st century education'. It is the extension of the inquiry-based approach to assessment that poses important challenges in this particular area (Griffin et al 2012).

Inquiry is a term widely used in everyday life as well as in education and other professional activities. It is sometimes equated with research, investigation, or 'search for truth'. Within education, inquiry can be applied in all subject domains – such as history, geography, the arts, science, mathematics and technology – when questions are raised, evidence is gathered and possible explanations are considered. In each area different kinds of knowledge and understanding emerge. Inquiry is not a new concept in education. It has roots in the studies of Piaget (1929) and the insights of Dewey (1933) and Vygotsky (1978), among others, in the first half of the 20th century, which drew attention to the important role in their learning of children's curiosity, imagination and urge to interact and inquire.

What characterises inquiry in education is that students are taking an active part in developing their understanding and learning strategies. They do this by pursuing questions or addressing problems that engage their attention and thinking. They bring their existing experience and ideas to bear in tackling the new challenge and in doing so strengthen and extend their ideas and strategies. Because they collect information for themselves, they have the evidence for what works and what does not work in helping them to make sense of different aspects of the world around. As well as building understanding they are developing competences such as critical thinking, communication skills and ability to learn both independently and collaboratively (Duschl et al 2007).

Inquiry-based education is firmly rooted in what we know about students' learning (Pellegrino et al 2001; Gopnik et al 1999), for example that

- children are forming ideas about the world around them from birth and will use their own ideas in making sense of new events and phenomena they encounter;
- direct physical action on objects is important for early learning, gradually giving way to reasoning, first about real events and objects and then later about abstractions;
- students learn best through mental and physical activity, when they work things out through their own thinking in interaction with adults or other students, rather than receiving instruction and information to memorise;
- language, particularly discussion and interaction with others, has an important part in the development of reasoning skills and ideas;
- teachers have a key role in students' learning in promoting active rather than passive learning.

These characteristics are closely aligned to the conditions identified by the IB (2008) as enabling students' best learning. They justify the attention given to inquiry-based learning

and teaching in programmes such as those of the IB because, when carried out effectively, inquiry-based education can lead to the concepts, skills, attitudes, action and the development of personal attributes that are the goals of learning in the 21st century.

The caveat 'when carried out effectively' is an important one, since the popularity of inquiry-based education makes it vulnerable to some misunderstandings and to being applied to practices which fall short of intentions (see, for example, Minner et al 2010). For example, a common misconception is that inquiry means that students have to 'discover' everything for themselves and should not be given information by the teacher or obtain it from other sources. This assumes that students come to new experiences with open minds and develop their ideas by inductive reasoning about what they observe and find through their inquiries. The reality is that students come to new experiences not with empty minds, but with ideas already formed from earlier thinking and experiences, which they use to try to understand new events or phenomena (Driver 1983; Driver et al 1985; Gopnik et al 1999). If there is no evidence to support their ideas then they need access to alternative ideas to try, which may be suggested by other students, the teacher or other sources.

The reference to the ideas that students form through their experiences within and outside school indicates a view of learning described as constructivist, since learners are involved in constructing rather than just receiving knowledge. However, the additional emphasis in inquiry on students working with others, on communication and dialogue, indicates a socio-cultural constructivist perspective (Bliss 1993). In this view the focus is on understanding through 'making sense of new experience with others' rather than by working individually. Knowledge is constructed communally through social interaction and dialogue. Physical resources and language also have important roles, as James (2012) 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. (James 2012: 192-193)

So what would assessment based on a sociocultural constructivist perspective on learning look like? This is a key question that educators advocating or implementing inquiry-based education are currently facing.

## 1.3 Assessment within the PYP

### The nature of PYP assessment

The general approach in relation to 'assessing' within the PYP is expressed in the statement that 'a well-designed learning experience will provide data on students' knowledge, skills

and conceptual understanding, and is consequently a vehicle for summative or formative assessment.' Thus quality of assessment is dependent on the nature and quality of the learning experiences, and the 'criteria for effective assessments' bring together a range of qualities which apply equally to effective learning experiences for students (OECD 2013; Alexander 2010; Stiggins 2001).

Student assessment in the PYP is identified as a process which involves 'gathering and analysis of information' and identifies 'what students know, understand, can do, and feel at different stages in the learning process.' Assessment is described as being 'wholly internal' and having the 'prime objective' of providing feedback on the learning process (IB 2007). Thus the overwhelming emphasis is on formative assessment (assessment for learning), which is seen as linked or interwoven with teaching, providing regular and frequent feedback. Summative assessment (assessment of learning) features strongly at the end of each unit and in the final year of the programme when the collaborative PYP Exhibition is assessed. Interestingly, summative assessment is considered to embody an element of formative assessment, since it is described as a process which enables students to show what they can do and 'informs and improves student learning'. [See section 2.2 for the origin of terms used to describe the formative and summative purposes of assessment.]

PYP schools are provided with information and guidance on assessment practice which they are expected to use in planning and implementing the assessment component of the curriculum. In their curriculum planning teachers are advised to consider the assessment of understanding, knowledge, skills and attitudes. This requires that the goals of activities (inquiries) are clearly articulated and ways of assessing achievement in relation to them are identified. The guidance on planning assessment urges teachers to be constantly looking for evidence relating to the criteria and to involve students where possible in planning assessment tasks.

It is recognised that, as in the case of the content and pedagogical components of the curriculum, there are likely to be national, state or regional requirements which need to be accommodated in schools' assessment programmes.

### Guidance on recording

Schools are advised to document students' learning using a range of methods. The guidance on 'recording' includes several ways of keeping records of what students have learned. These are described in terms of combinations of strategies (approaches to gathering data) and tools (particular instruments for collecting and recording data). The list of strategies for gathering data includes some referring to the process of data collection (e.g. 'observation') and some to the context or type of task in which students are engaged when data are collected (e.g. 'performance assessments'). Tools identified include the open recording of observations of students and the use of closed checklists of specific elements that ought to be present.

The IB does not provide external moderation or tests. It is acknowledged, however, that some IB schools in certain countries are required to use standardised achievement tests or national tests. For those where such testing is not a requirement, but undertaken voluntarily, schools are advised to 'carefully consider' points relating to reliability, validity (but without using these terms) and impact.

The use of portfolios of evidence of students' achievements is suggested as a means of documenting 'both the process of learning and the product, including images and evidence of students in the process of constructing meaning'. The potential of portfolios as a tool for assessment is also suggested. The use of ICT in creating and updating a portfolio is not mentioned. But ICT is used by students to support inquiry and in assessment; for example, in a unit relating to 'Sharing the planet' where an assessment task was to create a poster showing understanding of the factors that can lead to the extinction of plants and animals, the resulting posters were uploaded to the class blog with each student then asked to peer-assess the work of three other students.

### Reporting requirements

The PYP reporting requirements emphasise the importance of feedback. They treat together the feedback to learners and parents about students' performance and the feedback into the planning of teaching from the assessment of students. In relation to reporting on students' learning, the IB requires schools to report on the attributes in the learner profile (striving to be inquirers, knowledgeable, thinkers, etc.) but not necessarily at each reporting point. The purpose of doing this occasionally is to show that the school takes the development of these attributes seriously.

Reporting takes the form of written reports and conferences. Conferences are face-to-face events involving teacher, student and parents in various combinations for different purposes (e.g. students taking the lead in sharing their learning with their parents serves the purpose of encouraging them to take responsibility for their learning). Written reports are records for the students, parents and the schools which 'reinforce the underlying values of the programme'. In order to do this, any locally required report forms should be supplemented to reflect the aims of the PYP assessment model.

### The PYP Exhibition

The exhibition is a key part of the PYP assessment model, especially for students who will be continuing into the MYP. It takes place in the final year of the programme (normally the 6$^{th}$ year, but can be earlier in different school organisations), replacing one of the six units normally undertaken each year. The exhibition has a number of purposes, relating to students taking responsibility for their learning, working collaboratively and also independently. To this end, it is intended to address aspects of all transdisciplinary themes, use and develop all the transdisciplinary skills, and concern a topic related to real-life issues.

The *Exhibition Guidelines* for teachers follow the same pattern as the guidelines for planning units of inquiry, with added notes which largely concern matters of assessment. These emphasise, for example, that 'There should be assessment of each individual student's contribution to and understanding of the exhibition'. An 'exhibition reflection tool' is provided which can be used to evaluate various aspects of the exhibition in relation to practices set out in the relevant Standard of the PYP programme.

There are detailed guidelines for students which teachers make available to them in age-appropriate forms. These guidelines cover the discussion within the group working together on possible real-life problems or issues to consider as a topic for the extended inquiry, hints on planning and gathering materials, and the importance of keeping a record of sources, decisions and reflections in an individual journal. Students are advised on planning and preparing a presentation of the exhibition.

Assessment of group and individual performance is carried out by the students' own teachers, who will have been involved in choosing the topic of inquiry for the exhibition. Teachers are expected to decide what elements are to be assessed and to develop their own rubrics for assessment. The summative assessment is based on observation of participation and action taken during the exhibition, the presentation and response to questioning and individual students' journals.

### Action by students in the PYP

The expectation of students taking some action as a result of reflection on their inquiries is a unique dimension of the PYP. It lays a foundation for 'community and service' in the MYP and 'creativity, action and service' in the DP (IB 2008). It is a means for students to show that they have linked their classroom learning to real life situations and are developing responsible attitudes towards the physical environment and to the community within and beyond school. Teachers are asked to record in their reflection at the end of each unit and of the exhibition what student-initiated actions arose from the learning. For PYP students such actions will vary considerably according to age and experience. For example, after a unit on human migration some 10 year olds researched their family histories while others collected writing materials for schools in need of such resources in Africa; a 5-year old saved water run off before a shower to water his garden in order not to waste it.

## 1.4 The contents of this report

Before presenting and reflecting on the findings from the literature review we begin, in section 2, by offering a clarification of terms, since there is evidence that confusion persists, within IB documentation and within the academic literature at large, about the meaning of some quite critical assessment terminology.

In section 3 we overview principles of and approaches to assessment, moving on in sections 4 and 5, respectively, to consider the essentials of formative assessment (assessment *for* learning, which we argue in section 2 subsumes assessment *as* learning) and summative

assessment (assessment *of* learning), along with examples of practice as described in the literature and any well-founded evidence of impact on learning or learning motivation. Within sections 3, 4 and 5 we identify implications for PYP assessment practice. In section 6 we reflect further on formative and summative assessment practice within the PYP, and offer a list of what we consider to be the key implications at this point.

# 2. Key assessment terminology

*Over the past 25 years or so, the desire to set demanding standards of achievement for students and to know whether these are being achieved has led to a considerable increase in attention to assessment. At the same time new goals of learning have been identified in response to the changes following from the applications of technology in daily life and in education. It is important for changes in assessment to match these developments if it is to be useful in providing information to educational policy makers as well as to teachers, students, parents and others, and also to improve learning. As part of the development of assessment practices to support these functions, understanding of the process of assessment, of its strengths and limitations, has also developed. Clarity in the use of words is essential for the discussion, and so in this section we review current understanding of the meaning of assessment and of other terms which are used in describing its purposes and characteristics.*

## 2.1 Assessment, evaluation and appraisal

Assessment, as used in this report, is distinguished from evaluation and appraisal in the same manner as set out by the OECD:

> 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. (Nusche et al 2012: 24)

Whilst clear in the English language, the distinction between assessment and evaluation is less easily made in many other languages, such as Spanish and French, where there is only one word used for both terms. In these cases it is important to qualify the term to indicate whether it refers to students or to programmes, institutions or policies. It is also important to recognise that, although student assessment will have a role in programme and school evaluation, there will be many other kinds of information that need to be taken into account. Care must be taken not to equate programme or school evaluation with student assessment data alone, for to do so can lead to unfair judgements of the effectiveness of a programme or school. Evaluation of programmes and schools should take account of context and background variables as well as student achievement.

There is sometimes confusion in the discourse around assessment concerning whether it refers to a process or the product of a process. In the OECD description above the statement that assessment refers to judgements suggests an outcome, but more commonly it is defined as a process, as in the succinct and widely quoted statement of Popham (2000) that assessment is "a process by which educators use students' responses to specially created or naturally occurring stimuli to draw inferences about the students' knowledge and skills". In this report we use the term assessment to mean a process and refer to the outcome of the process as assessment results.

## Definitions of assessment

The conception of assessment as a process of drawing inferences from data is built upon by Pellegrino et al (2001: 42) in describing assessment of all kinds as 'reasoning from evidence' (a phrase also used by Mislevy 1996). This process of reasoning is described by Pellegrino and colleagues in terms of three key elements:

- cognition (a model or set of beliefs about how students learn)
- observation (the tasks to which students respond in order to show their learning)
- interpretation (the methods and tools used in the reasoning that turns data into information).

These elements are represented as the corners of a triangle in which each is connected to the other two as in figure 2.1 (which is an adaptation of Pellegrino and colleagues' original assessment triangle). The connection between the observed task-based performances (observations) and the judgements made of them (interpretation) represents the dependence of what is assessed on both. This is particularly clear in the case of items in a test: what a simple test item assesses will depend as much on the mark scheme (rubric) as on the question, but it applies equally to the results of interpreting any form of data. The links between 'observation' and 'interpretation' and 'cognition' indicate that both the tasks and the interpretation of the responses are judgements which depend on values and on the view of what learning means.

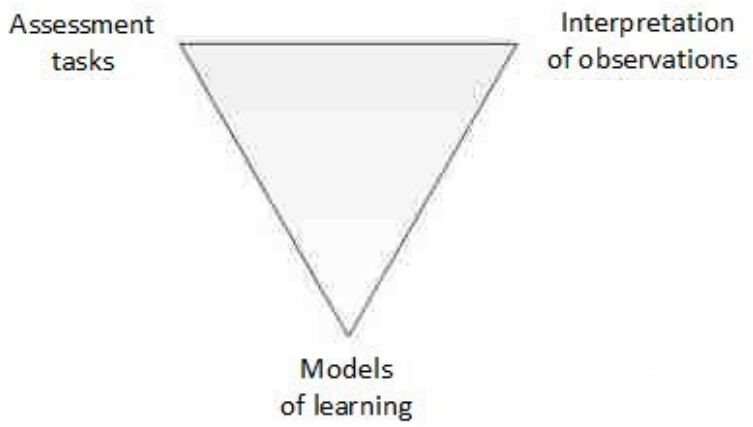


Figure 2.1 *The Assessment Triangle (adapted from Pellegrino et al 2001)*

Beliefs about how learning takes place will influence the kind of data sought and collected and how it is turned into information about the knowledge and skills being assessed. It underlines the need for the tasks and situations that enable students' performances to be observed to be consistent with the model of learning. Similarly, the methods of interpretation through which inferences are drawn must also reflect the model of how students learn. If it does not, then 'the meaningfulness of inferences drawn from the assessment will be compromised' (Pellegrino et al 2001: 54). These are matters which affect the validity of assessment, to which we return later.

Other definitions spell out the processes involved in assessment in more operational detail. For example, Harlen (2013) describes assessment as the generation and collection of data, the interpretation of the data to produce a judgement, and the communication and use of the judgement. The inclusion of 'generation' here is a reminder that decisions are made about what will be taken as evidence of learning and achievement, for there has to be a selection from the innumerable possibilities in students' behaviour. Normally the goals of students' activities will provide a focus, but then there are also numerous decisions (influenced, as indicated in the assessment triangle, by the implicit or explicit view of learning) to be made about:

- the kind of activity in which students are engaged (e.g. special task or regular work);
- who will collect data from that activity (e.g. teacher, students or an external agent);
- how the data will be collected (e.g. by observation or as writing, drawing or other artefacts);
- how the data will be interpreted (e.g. by reference to norms or criteria);
- how the resulting judgement will be recorded and communicated (e.g. as a score, grade, or description).

## 2.2 Formative and summative assessment

According to Atkin and Black (2003), the idea of making a distinction between formative and summative purposes originated in the late 1960s (Tyler et al 1967) in the context of curriculum evaluation, where formative evaluation described the effort to improve a programme rather than to judge it definitively. The distinction was soon adopted in the context of student assessment as, for example, in the *Handbook of Formative and Summative Evaluation of Students' Learning* (Bloom et al 1971), to differentiate between classroom assessment as part of learning and teaching and assessment as judging learning outcomes. As the terms 'formative' and 'summative' appeared rather formal and academic, in communicating with teachers the alternative phrases 'assessment for learning' and 'assessment of learning' were also used by the Assessment Reform Group (ARG 1999) among others.

In theory it is possible to distinguish between purposes and uses of assessment, purposes being the reasons for conducting the assessment and uses being what is actually done with the assessment results (Harlen 2012). A common view of purposes, represented by the discussion in Pellegrino et al (2001), is that there are three main purposes of assessment:

- Assessment to assist learning – formative assessment
- Assessment of individual students' achievement – summative assessment
- Assessment to evaluate programmes.

Newton (2007) points out, however, that whilst the first of these is indeed a purpose, the second is not a purpose but a statement of the subject of the assessment, which could have a number of purposes and uses. In the case of formative assessment there is only one main use of the information, which is to help learning. If it is not used for this purpose then the process cannot be described as formative assessment (or assessment *for* learning). By

contrast, the data from summative assessment of students' achievement (or assessment *of* learning) 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 (Newton 2012). Newton (2010) listed 18 uses, including monitoring students' progress over time, placing students in teaching groups, diagnosing students' difficulties, reporting to parents, guiding future educational and employment decisions, etc. The uses also include evaluation of programmes and institutions. Thus the third purpose above can be regarded as a use of student achievement data rather than a different purpose. This suggests that the distinction between purpose and use becomes difficult to sustain. So it is helpful to follow the decision by Mansell et al (2009) to simplify this discussion by referring to uses in three broad categories:

1. The use of assessment to help build pupils' understanding
2. The use of assessment to provide information on pupils' achievements to … parents, to further and higher education institutions and employers
3. The use of assessment data to hold individuals and institutions to account.

In this review we deal only with the first two of these uses.

Assessment carried out as part of teaching and learning for the purpose of helping learning is described as 'formative' assessment, or assessment *for* learning. It involves teachers and students in using evidence of learning as it takes place to feed back into decisions about how to help progress towards lesson or unit goals. Key features of assessment *for* learning are listed in Box 2.1.

*Box 2.1 Key features of formative assessment (assessment for learning)*

- Feedback to the students that provides advice on how to improve or move forward, and avoids making comparisons with other students.
- Students understanding the goals of their work and having a grasp of what is good quality work.
- Students being involved in self-assessment so that they take part in identifying what they need to do to improve or move forward.
- Students engaged in expressing and communicating their understandings and skills, initiated by teachers' open and person-centred questions.
- 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. (Harlen 2007: 119)

A questionnaire survey of primary and middle school teachers (Sach 2012) found that the perceptions of formative assessment were largely in agreement with these statements, with some variation according to school phase and length of service.

Assessment carried out for the purpose of summarising and reporting students' achievement at a particular time, such as the end of a course or programme of study, is described as 'summative' assessment, or assessment *of* learning. It may or may not have some impact on learning and the outcome may or may not be used in later decisions about teaching, but such impact is not its main rationale. Key features of assessment *of* learning are listed in Box 2.2.

*Box 2.2 Key characteristics of summative assessment (assessment of learning)*

- Relates to achievement of broad goals expressed in general terms rather than the goals of particular learning activities.
- Results reported at certain times not as an everyday part of learning.
- Uses evidence obtained when students are involved in special tasks or tests as part of, or in addition to, regular work.
- May be based on teachers' judgements, tests or a combination of these.
- Involves judging the achievement of all students against the same criteria or mark scheme.
- Requires some measures to assure reliability.
- Typically provides limited, if any, opportunities for student self-assessment.

It is important to note that the *process* of assessment is the same in both cases, and many of the available tools and strategies for carrying out assessment can be used in either context. The adjectives 'formative' and 'summative' describe differences in *purpose* not in form. In theory a distinction can be made between using assessment to help learning and using it to report on what has been learned, but in practice the distinction is not a sharp one.

The lists of characteristics of formative and summative assessment in Boxes 2.1 and 2.2 emphasise the differences between the two assessment *purposes*, differences which should be kept very clearly in mind, especially when both are carried out by teachers. It is too often assumed that all assessment by teachers is formative or that assessment carried out frequently in whatever way is formative. Unless the assessment is used to help the ongoing learning, this is not the case and the true potential value of formative assessment will not be realised. The research by Phelan et al (2011) illustrates the interpretation of formative assessment as a series of tests to check understanding. The results showed that the intervention, which also included professional development for the teachers, led to improved performance. The improvement was greater for the higher scoring students, in contrast with the findings reported by Black and Wiliam (1998a) which showed the greatest gains were for the lower achievers.

A further distinction is that summative assessment is a necessary part of educational practices because reports on students' learning have to be made for a number of reasons and records have to be kept; there is no choice about needing some summary of students' learning. By contrast, formative assessment can be considered as voluntary in the sense that teachers

decide the extent to which it is included in their practice. Moreover, formative assessment, being integral to teaching, is invisible; summative assessment is generally very visible. Research (e.g. Pollard et al 2000 in England) has shown that when summative assessment has a very high profile it can dominate the classroom ethos and risks driving out the formative use of assessment. This results in the achievement of a grade or level becoming more important to students and teachers than effort and evidence of progress. So, since summative assessment is unavoidable, it is important to find ways in which it can be carried out and used without having a negative impact on formative assessment.

The use of summative assessment to help learning was one of four practices, reported by Black et al 2003, that teachers found were effective ways of implementing formative assessment. However, the essentially summary nature of summative assessment means that it is generally not sufficiently detailed to feed back into specific activities and at best enables students to revisit unsuccessful parts of their work.

Since formative assessment is in the hands of the teacher it follows that in this approach the evidence to be used in summative assessment will have been collected by teachers. This raises two potential problems: first that the evidence gathered is dependent on the learning experiences provided; and second that it depends on the ability of the teacher to collect evidence. Moving from detailed evidence gathered during day-to-day learning tasks to a summary of achievement requires assurances about the breadth of the curriculum experienced by the students, so that the evidence used is valid and adequately reflects the learning goals, and that the judgements are reliable.

### A note about 'assessment *as* learning'

Some authors on assessment have proposed the concept of 'assessment *as* learning' as a separate purpose of assessment. The main source of this idea is the work of Lorna Earl and her colleagues in Canada. They propose that

> 'Assessment as learning' is a process of developing and supporting metacognition for students. Assessment as learning focuses on the role of students as the critical connector between assessment and learning. When students are active, engaged and critical assessors, they make sense of information, relate it to prior knowledge and use it for the new learning. This is the regulatory process in metacognition. It occurs when students monitor their own learning and use the feedback from this monitoring to make adjustments, adaptations and even major changes in what they understand. (WNCP 2006)

The authors also acknowledge that what is described here is included in the description of formative assessment above (see Box 2.1) and that many authors describe classroom assessment as having only two purposes: assessment for learning and assessment of learning. The identification of student self-assessment as a concept separate from assessment for learning is not widely reflected in the current literature. Moreover, it has also been suggested that conceiving assessment as learning relates to a less positive relationship between assessment and learning in which assessment practices come to

dominate learning experiences (Torrance 2007). In this report we consider the student role in using assessment to help their learning (metacognition) as part of formative assessment.

## 2.3 Validity, reliability, dependability

These are properties of assessment tools and processes that influence decisions about the selection of the methodology which would best serve the purpose in a particular case. For example, if the assessment result is to be used to select individual students for a particular course or award, then the methods chosen needs to have high reliability (repeatability). On the other hand, for an assessment used by teachers and students to identify and help progress or diagnose a difficulty, validity is more important than reliability. As explained later, these properties interact and making changes that effect one has implications for the other.

Validity refers to whether the assessment process provides the kind of information that is intended and serves the purpose of the assessment. Reliability refers to how much one can trust the results of the process; that they are as precise and replicable as possible. Figure 2.2 illustrates these two concepts, which, though considered separately here, are to some extent interdependent in practice. The concept of dependability, or assessment quality, is sometimes found useful in describing the combination of validity and reliability.

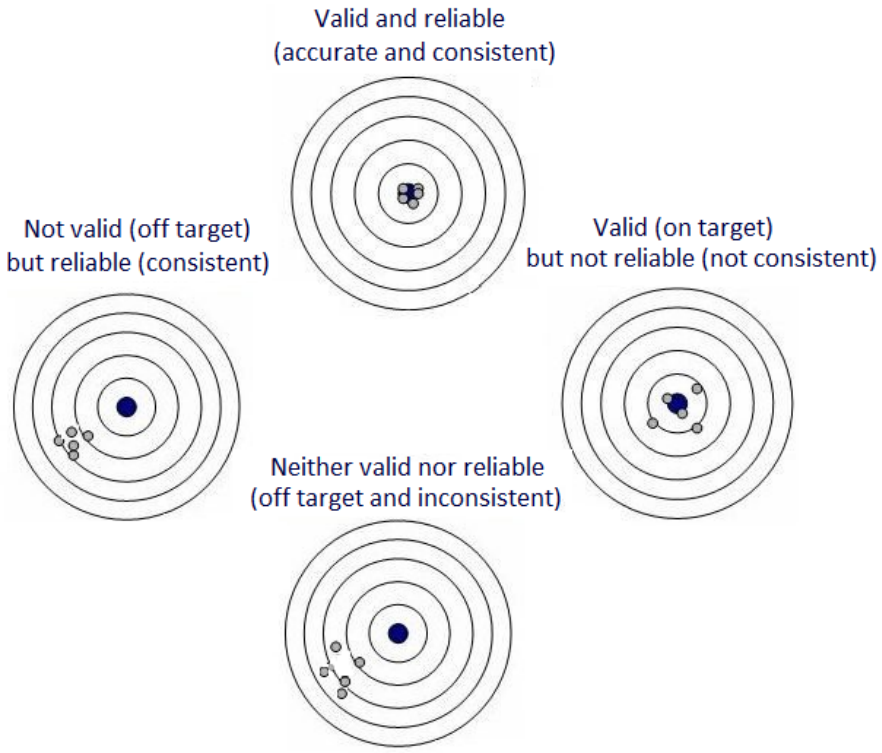


Figure 2.2 *The concepts of validity and reliability (adapted from Johnson 2012)*

### Validity

There are various ways of judging the validity of an assessment instrument, such as a test or holistic performance task, or of a set of procedures, such as a scheme for gathering observational data during regular classroom work. One way is just to look to see if it appears

to require the use of the knowledge or skill that it is intended to assess. This is described as 'face validity'. A more formal version of this is 'content validity', where there is some analysis of what content is covered by the assessment and what is found is checked to see if it reflects the range of content intended to be assessed. 'Construct validity' is a broader concept relating to the full range of skills, knowledge or attitudes – the constructs – that are in principle being assessed. Johnson (2012: 60) cites 'reading comprehension', 'numeracy', 'sense of community', 'attitude to mathematics', as examples of constructs.

An important requirement of an assessment is that it 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. The amount of reading required by a question in a test where reading ability is not the construct being assessed is a common problem. Often the attempt to set a question or task in a real context in order to engage students' interest, or to see if they can transfer learning into other contexts, extends the reading demand. What is then assessed is a combination of reading skills and the intended construct, and validity as a measure of the intended construct is consequently reduced. The assessment is said to suffer from 'construct irrelevance' since the result will depend on other things than the intended construct.

Another form of validity, consequential validity, is not a property of the assessment instrument or procedure itself, but is a comment on how appropriate the assessment results are for the uses to which they are put. The validity of an assessment application is reduced if inferences drawn on the basis of the results are not justified, either because other things than the intended construct are being assessed ('construct irrelevance') or because it does not include some aspects of the construct. For example, a test of arithmetic may be perfectly valid as a test of arithmetic but not valid if used to make judgements about mathematical ability more generally. This is described as 'construct under-representation'. The message is 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. (Messick 1989:13)

Note that consequential validity is not to do with the quality of the assessment process itself or of its immediate results. It is rather to do with the appropriateness of the uses made of the assessment results for purposes other than, and including, those for which the original assessment was designed and conducted.

### Reliability

The reliability of an assessment refers to the extent to which the results can be said to be of acceptable consistency and precision for a particular use. There are many factors that can reduce the reliability of an assessment. Reliability is reduced if, for instance, the results are influenced by the particular set of items or tasks a student is required to attempt. It will be

reduced also if the outcomes are dependent on who conducts the assessment (which teacher or oral examiner), and on who rates the students' assessment performances (which marker or observer). The particular occasion on which the assessment takes place, and the circumstances under which it is undertaken, can also affect the assessment outcome and contribute to reduced reliability. A critical concept here is the potential *variability* in the outcomes of assessment.

Reliability is defined as, and, where possible, estimated quantitatively by, the extent to which the assessment, if repeated, would give the same result: replicability is the key property here. The most common quantitative indicators of reliability are a coefficient of correlation, taking values between 0 (totally unreliable) and 1 (totally reliable) and a confidence interval. When teachers rate students' performances holistically then the reliability coefficient would be the correlation between the judgements of teachers made independently. In the case of multi-item tests a very commonly used indicator (because it is available as a push-button option in most statistical software packages, including SPSS) is Cronbach's alpha coefficient, an index developed more than half a century ago and based on the average inter-item correlation in the test.

The weakness in these correlation-based methods is that they address just one single source of unreliability at a time (they are single-variable approaches), whereas in reality there are often several sources of unreliability at play simultaneously. For example, in a writing assessment there are at least two important contributors to unreliability: one is the topic of the writing assignment, another is the individual who marks the writing produced. In a group discussion task there are at least three important sources of unreliability in assessment: the topic given to the students to discuss, the observer who rates the performance of the group or of individual students during the discussion (or later on video), and the composition of the discussion group itself. Had the student been offered a different topic to write about, or a different topic to discuss within a group, would the resulting assessment of writing or discussion skills have been the same? Had a different teacher rated the student's writing or the student's contribution to the group discussion, would the resulting individual assessment have been the same? Had the student been a member of a different group for the same discussion would the student's performance, and the group performance, have differed?

The appropriate way to estimate the reliability of assessment results in these authentic situations is to produce a generalizability coefficient, or, better, a confidence interval around the assessment score or judgement, which takes into account as many of the error-contributing influences as possible simultaneously (for more information see Johnson 2012, chapter 4).

### Dependability

Validity is paramount in any kind of assessment; the assessment must provide information about all relevant goals and attributes related to learning. In formative assessment reliability

is less important because of the ongoing nature of the process. The information is used to inform teaching in the situations in which it is gathered. Thus there is always quick feedback for the teacher and any misjudged intervention can be corrected. Thus considerations of reliability do not need to impact heavily on validity. This is not to say that teachers do not need to consider how well they gather and interpret evidence, but they do not need to be overly concerned about reliability as is necessary when judging it in terms of grades or levels.

Formative assessment is concerned with the future not with judgements about the past. As Black and Wiliam (2012: 260) note: 'formative assessment is reliable to the extent that the assessment processes being used generate evidence that consistently lead to better, or better founded, decisions'. For summative assessment, however, reliability is very important since its purpose is to provide information about where students have reached in their learning that parents, other teachers and educational managers and policy makers can depend upon. So attention has to be given to increasing reliability as far as possible without endangering validity.

The reference to 'increasing reliability as far as possible without endangering validity' hints that these aspects of an assessment process are not independent of one another. In any summative assessment there is a limit to the extent to which both reliability and validity can be optimised. The argument that leads to this conclusion is that if reliability is low there must be some unintended factors influencing the result and therefore what is being assessed is uncertain. To increase reliability it is necessary to reduce the unwanted influences, marking variability and task influences being two of the most important. This most often leads to using tasks and methods where outcomes can be consistently judged, meaning more focus on factual knowledge, where there is a clear right answer. But if the purpose is to assess skills and understanding, which are best assessed when students are asked to generate rather than select answers, using such closed tasks would reduce assessment validity. On the other hand, to increase validity by including more open-ended tasks could reduce reliability because the marking or scoring would be less clear cut, resulting in marker variability; it could also lead to the assessment becoming time-consuming and even unmanageable because several different tasks should probably also be used. Thus there is a trade-off between reliability and validity; increasing one can decrease the other when the time and budget for assessment are constrained.

The term 'dependability' was introduced into educational assessment in the 1950s by behavioural science researchers in the US to acknowledge the interdependence of the concepts of validity and reliability (see in particular the seminal text of Cronbach et al 1972). Put simply, an assessment may have low dependability either because, although having high construct validity (such as the performance of a skill), it cannot be reliably judged, or because, although reliably judged (as in answers to a multiple choice test), what is assessed is a poor reflection of the construct.

It is clear from this conception of dependability as the 'intersection of reliability and validity' (Wiliam 1993), that it has no meaningful numerical value. However, it can be a convenient term for the overall quality and usefulness of an assessment. In general, increasing the dependability of an assessment involves searching for optimum reliability whilst ensuring acceptable levels of validity, however these might be defined. This prioritises validity, but the trade-off has to take account of the intended use of the assessment results; some uses require greater efforts to improve reliability than others – which leads to matters of manageability (i.e. the cost in terms of time and resources of operating an assessment process of the required validity and reliability).

## 2.4 Manageability

Stobart (2008) points out that there are several factors that 'pull against' the development of assessment procedures that provide the most dependable information for a particular purpose. One of these is indeed 'the pull of manageability' which he describes as the search for simpler and more cost-effective assessments. This refers to resources of time, expertise and cost to those involved in creating and using the assessment materials and procedures. 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 by external examiners; 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 section 4, 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 assessment information to identify next steps in learning. These costs, however, are integral to efforts to improve teaching and 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 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 and scoring tests are added to the indirect costs of class time taken up by preparing for and taking external tests and examinations, the total can amount to a significant proportion of the education budget (Harlen 2007: 61-62). It certainly constitutes a case for considering the balance between costs and benefits in deciding the methods to be used for summative assessment.

But that there is a trade-off in any assessment has to be acknowledged as an unavoidable feature of a strategy or tool. Taking cognisance of this point is essential to understanding assessment results as approximations and the consequent limitations on the inferences that can be drawn from them.

## 2.5 Norm-referenced, criterion-referenced and diagnostic tests

A standardised, or norm-referenced, test comprises items and tasks created by an external agency through a process in which the test is trialled using a large representative sample of the appropriate population, so that individual students' scores can be expressed in terms of comparisons with the 'norms' for that population. The results will indicate whether an individual student's performance is above or below average but not necessarily what that student knows and 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, irrespective of the performance of other students. The items – the test questions – will be chosen for their relevance to the curriculum so that the results can be used in establishing, not how particular students compare with others, but how their performances compare with the intended performance. At the same time, the target level of performance is set by reference to what educational professionals think 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 (Black 1998).

Diagnostic tests are specially designed to provide information about the strengths and weaknesses of individual test-takers. Typically such tests are narrowly focused, and even then they need to contain enough test items so that areas of knowledge/skill weakness can be identified reliably.

# 3. Assessment principles and approaches

*This section focuses on principles of assessment, and on approaches to the assessment of conceptual understanding, knowledge, skills and attitudes. Common strategies and tools for assessment are also considered here.*

## 3.1 Principles of assessment

All statements about the aims and outcomes of education are based on what is considered to be of value and worthwhile to learn. This is of particular relevance in the case of assessment, since those aspects of learning that can be included in any form of assessment are inevitably only a sample of what could potentially be included. Decisions have to be taken about what to include and these need to be justified. The validity of the resulting assessment is highly dependent on this process. So it is important to make explicit the values underpinning the decisions about what to assess, how, and for what purposes. The vehicle for this is generally to set out the principles, or the aspirations, for the assessment. Such principles embody the view of what is 'good' or 'quality' assessment.

At the most general level the OECD has set out the key policy directions identified from its reviews of evaluation and assessment in education in its member countries, summarised in *Synergies for Better Learning: an international perspective on evaluation and assessment* (2013). The points of most relevance to student assessment are:

- Integrate student assessment and school evaluation in a single framework.
- Align assessment and evaluation with educational goals and learning objectives set out in the curriculum.
- Focus on improvement of classroom practices and build on teacher professionalism.
- Design the accountability use of evaluation and assessment in ways which minimise undesirable effects.
- Place the student at the centre, fostering engagement in learning through using formative assessment strategies.
- Use measures of performance broad enough to capture the whole range of student learning objectives.

At a more detailed level are four 'Guiding Principles' which Stiggins (2001: 17) identifies as representing 'important values that I personally have come to hold about classroom assessment'. They are:

1. *Students are the key assessment users*.

   Stiggins points out that although there are other users of assessment results it is only students who use the results to set expectations of themselves and set their sights on what they can achieve based on their success in classroom assessment.

2. *Clear and appropriate targets are essential.*

   The point here is that good assessment depends on precise definition of what it means to do something well.

3. *Accurate assessment is a must.*

   Accuracy is important because inaccuracy places 'student academic well-being in jeopardy' (p22). Accuracy requires clarity of targets and of purpose and the choice of a method that is 'capable of reflecting the valued target' (ibid).

4. *Sound assessment must be accompanied by effective communication.*

   This does not necessarily mean communicating achievement in terms of scores. 'We can use words, pictures, illustrations, examples and many other means to convey this information' (p23).

Stiggins' list focuses on classroom assessment by teachers. The Assessment Reform Group (ARG) also produced a list of 'research-based principles to guide classroom practice', identifying 10 principles specific to assessment for learning, in Box 3.1.

*Box 3.1 10 Principles of Assessment for Learning (ARG 2002a)*

Assessment for learning should:

- be part of effective planning of teaching and learning;
- focus on how students learn;
- be recognised as central to classroom practice;
- be regarded as a key professional skill for teachers;
- be sensitive and constructive because any assessment has an emotional impact;
- take account of the importance of learner motivation;
- promote commitment to learning and a shared understanding of the criteria by which they are assessed;
- help learners receive constructive guidance about how to improve;
- develop learners' capacity for self-assessment so that they can become reflective and self-managing;
- recognise the full range of achievements of all learners

The ARG list emphasises that 'assessment processes are an essential part of everyday classroom practice and involve both teacher and learners in reflection, dialogue and decision making' (ARG 2002a), indicating a view of formative assessment that includes the role of students in self-assessment, knowing the goals of their work and the assessment criteria and taking part in decisions about how to improve and make progress.

Other lists suggest principles applicable to all assessment contexts and purposes. One of these, a widely quoted list of 10 principles of assessment, was identified in the course of a

project studying how to improve formative and summative assessment by teachers but nevertheless was considered to be applicable to all assessments. It was suggested that these principles (Box 3.2) should be used as standards to aim for in planning or evaluating assessment procedures or proposals.

*Box 3.2 Principles of assessment practice (Gardner et al 2010, p20)*

- Assessment of any kind should ultimately improve learning.
- Assessment methods should enable progress in all important learning goals to be facilitated and reported.
- Assessment procedures should include explicit processes to ensure that information is valid and as reliable as necessary for its purpose.
- Assessment should promote public understanding of learning goals relevant to students' current and future lives.
- Assessment of learning outcomes should be treated as approximations, subject to unavoidable errors.
- 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 achievements will be judged.
- Assessment methods should promote active engagement of students in their learning and its assessment.
- Assessment should enable and motivate students to show what they can do.
- Assessment should be based on information of different kinds, including students' self-assessments, to inform decisions about students' learning and achievements.
- Assessment methods should meet standards that reflect a broad consensus on quality at all levels from classroom practice to national policy.

The principles in Box 3.2 were a starting point for a group of primary school educators developing proposals for assessment in primary schools. Two sets of principles resulted (Box 3.3), one relating to the functions of assessment, the other to methods and practices.

*Box 3.3 Adaptation of the principles in Box 3.2 by primary school educators*

*In relation to its functions, assessment should:*

- ultimately improve pupils' learning;
- be aligned with the full range of learning objectives of the whole school curriculum;
- be an integral part of teaching that enables pupils to understand the purpose of their activities and to improve the quality of their work;
- combine qualitative and quantitative data of different kinds from a variety of individual and group learning activities including pupils' self-assessment, to inform decisions about pupils' learning and achievements;
- be understood as providing data about pupils' learning outcomes in the form of approximations and samples, subject to unavoidable variations.

*In relation to the methods and practices of assessment, these should:*

- promote the active engagement of pupils in their learning and its assessment, enabling and motivating them to show what they know and can do;
- include explicit processes to ensure that information is valid, reflecting all important learning goals, and is as reliable as necessary for its purpose;
- meet standards that reflect a broad consensus on quality at all levels from classroom practice to national policy;
- be realistic and manageable for pupils and teachers, with transparent time demands and requiring no more collection of pupils' work than is a normal part of teaching and learning.

### The PYP principles of assessment

The key principles of the IB assessment, which apply to all three programmes (PYP, MYP and DP), are stated as in Box 3.4.

*Box 3.4 IB principles of assessment*

- Assessment is integral to planning, teaching and learning.
- The assessment system and assessment practices are made clear to students and parents.
- There is a balance between formative and summative assessment.
- Opportunities for peer and self-assessment are planned for.
- Opportunities for students to reflect on their own learning are planned for.
- Students' current knowledge and experience are assessed before embarking on new learning.
- Students are provided with feedback as a basis for future learning.
- Reporting to parents is meaningful.
- Assessment data is analysed to provide information about the teaching and learning, and the needs of individual students.
- Assessment is used to evaluate the effectiveness of the curriculum.

By comparison with the lists above and particularly Box 3.2, some of the IB principles of assessment in Box 3.4 appear to concern operational features rather than quality expectations. Although echoing several of the policy directions for assessment and evaluation identified by the OECD (2013), these statements are expressed in almost value-free terms. For example, the first item in the list in Box 3.4 could refer to an assessment regime that is integral to planning, teaching and learning but yet is so rigid that it restricts the development of qualities in the IB learner profile. It does not make explicit the values underlying the IB profile that are undoubtedly intended to be read into the statements. This

intention could be signalled by a general statement prefacing the assessment principles or by qualifying each statement appropriately.

Although principles are intended to be general statements conveying values rather than evaluation criteria, if they are to have a role in guiding decisions about assessment policy and practice it has to be possible to know when they apply. It is, for instance, hard to decide when policy and practice match a statement such as 'There is a balance between formative and summative assessment'. Similarly 'Assessment is used to evaluate the effectiveness of the curriculum' could mean using the result of tests to decide how effective the curriculum is, which is very unlikely to be the intention and certainly not reflected in the description of curriculum evaluation in the self-study report or other programme evaluation procedures. In order for the principles to have a useful role in signalling the assessment philosophy of the IB it would be helpful for the statements to be carefully worded and accompanied by some discussion of each as it applies to the PYP.

## 3.2 Approaches to assessment

### Assessing conceptual understanding, knowledge, skills and attitudes

In order to gather data about the extent to which students are grasping concepts, acquiring knowledge, and developing skills and attitudes, it is necessary for them to be engaged in tasks in situations where the concepts, knowledge, skills and attitudes in question are required. These situations might be found in the course of regular classroom work or particular tasks might be designed to elicit evidence of understanding or the use of skills or indications of willingness to act in certain ways. Such tasks could be embedded in regular work or introduced as separate and special events, as in the case of tests. Examples of a range of classroom tasks designed to assess students' learning in English are described in Keogh et al (2008) and in science by Naylor et al (2004).

In the case of assessment of the development of concepts, understanding is shown in the ability to organise knowledge, to relate it actively to new and to past experience, forming big ideas, much in the way that distinguishes 'experts' from 'novices' (Bransford et al 2000).

> 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. (Pellegrino et al 2001: 72-73)

Big ideas are ones that can be applied in different contexts; they enable learners to understand a wide range of phenomena by identifying the essential links ('meaningful patterns') between different situations without being diverted by superficial features. Merely memorising facts or a fixed set of procedures does not support this ability to apply learning to contexts beyond the ones in which it was learned. Facts are only as useful as the

links that can be made between them. Understanding is shown when knowledge can be applied to new situations and can be used in problem-solving and decision-making.

Being able to apply a concept in a new situation indicates that some links have been formed between the new situation and the one in which it was learned, but there is clearly a limit to how far beyond current experience primary school students are able to apply their learning. This has implications for the kind of activity in which students are engaged so that conceptual understanding can be validly assessed. The new situation has to be familiar enough for links to be made: not so familiar that the task only requires recall rather than application, and not so unfamiliar than students cannot relate their existing knowledge to it. In addition, the task should be engaging to students taking into account their age and experience.

The familiarity of the tasks on which students are engaged when assessed is also a relevant factor to be considered in the assessment of skills. All skills have to be used in a context and in relation to some content. All tasks have to be about *something*; observation has to be about certain objects or events; reasoning and interpreting will be in the context of using evidence about some phenomenon or situation. The nature of this subject matter makes a difference to whether skills are used. Students might be able to plan an appropriate inquiry about a situation where they have some knowledge of what are likely to be the relevant factors to pursue, but fail to do this if the subject matter is unfamiliar. This has important consequences for assessment. The choice of subject has an influence and this means that assessment using only one situation or subject will likely give a non-valid and unreliable result. The best approach is to collect data across a range of situations in which the skills can be used.

An attitude refers to 'a learned predisposition to respond in a consistent favourable or unfavourable manner with respect to a given object' (Anderson and Bourke 2002: 31). Attitudes are an important, but not the only, part of the aspect of the affective outcomes of experience. Stiggins (2001: 345), quoting Anderson and Bourke, lists eight kinds of affect of relevance to schooling. These include attitudes, locus of control, self-efficacy, interests and academic aspirations. Several of these are overlapping concepts, all contributing to motivation for learning (ARG 2002b).

Attitudes concern not what students know or can do but what they feel about themselves, the tasks they undertake, other people and the world around. However, the assessment of attitudes is not a regular part of some assessment programmes for two main reasons: firstly, dependable measurement of attitudes is problematic, and, secondly, the view that attitude assessment ought not to be attempted in any case on ethical grounds. In the first of these, the argument goes that since attitudes concern feelings – and there can be no correct or incorrect responses, no judgements to be made, no marks or scores that can be given – they cannot be measured in the same way as cognitive outcomes. We also do not know how stable attitudes are in individuals, or how demonstration of attitudes might be affected by

the circumstances of the assessment. The second argument concerns a principled objection arising from the view that the teaching of personal dispositions and their assessment is a matter for parents rather than the school. However, most would agree with Griffith (2008) and with Stiggins (2001) that developing students' dispositions, particularly to learning, is a key responsibility of education.

> Indeed, we can do great harm if school leaves students feeling as though they are incapable of learning. Regardless of their actual ability to learn, if they don't perceive themselves as in charge of their own academic well-being, they will not be predisposed to become the lifelong learners they will need to be in the twenty-first century. Attitudes about writing, reading, and other academic domains, as well as academic self-content, are important targets of classroom instruction. (Stiggins 2001: 339)

This view has support from neuroscience which highlights the importance of emotions in learning. Studies of the brain reported by OECD/CERI (2008) have shown 'how negative emotions block learning', and that although some level of stress is essential for optimal response to challenges 'beyond this modicum it activates response in the brain associated with flight and survival and inhibits those responsible for analytical capacity' (OECD/CERI 2008: 4).

In order to ensure the most effective operation of students' brains, then, teachers need to know how students are feeling about their experiences and whether the affective goals of teaching and learning are being achieved so that action can be taken if necessary. The methods that can be used to access students' emotional response to their work are discussed in the next section.

## 3.3 Questioning, observing and evaluating products

The principal methods teachers use in the classroom to assess learning, and which are used in assessment more generally, are:

- questioning (including through dialogue)
- observing
- evaluating products

### Questioning

Questioning is a natural component in interactive teaching and learning. Questions can be short and direct, as in 'What is 10 multiplied by 5?', 'What is the capital city of Brazil?' or 'What does the price list say an ice cream costs?'. They can also be more 'open' and searching, such as 'Why do you think Jasper was sad when he saw the boat?'. There is a place for direct questioning in teaching and learning to assess factual knowledge and, to some extent, conceptual understanding. But it is widely acknowledged that open questions are the preferred form in formative assessment contexts, where the teacher can use them to explore learners' knowledge acquisition and conceptual development (see section 4 for a full discussion). A 'question' can be modified into an instruction and have the same power to elicit evidence of knowledge and conceptual understanding: an example would be 'Mark the place on the map where the new school is to be located'. Or it might become a

combination of question and further request, such as 'What did the Romans use to make foundations for their roads' followed by 'Explain why they made foundations that way…'.

The same kinds of question can equally be used in formal summative tests, where they become 'test items' by virtue of having an associated criterion-based mark scheme, which identifies the particular piece of knowledge or conceptual understanding the item has been designed to elicit. A difference between questions posed by the teacher informally in the classroom and those same questions presented as test items in a summative test can be in the form of student response required. In the case of teacher-posed questions the response from the student is usually a constructed response, i.e. a response produced without the benefit of cues. 'Constructed-response' items also feature in summative tests, but so, too, do 'objective' items, or 'select' item types, in which the student is presented with a list of possible answer options from which to choose the correct one(s): multiple-choice questions will be the most well-known format (for useful guides on item types and their development see Johnson 2012, chapter 3, and SQA 2009).

In formal tests constructed-response items can sometimes give misleading results, since students' written language skills, if poor, can interfere with their ability to provide good evidence of whether or not they have the knowledge and understanding that is in principle being assessed. Open questions also tend to be left unattempted by more students than do closed questions, because they do not have an answer or are not sufficiently motivated to make the effort to provide it, or simply because they have succumbed to test fatigue. In consequence, the legitimacy of such an item as a valid assessment tool must sometimes be in doubt.

The range of methods used to assess attitudes also includes written questions (questionnaires) and oral questions (interviews). Written forms often comprise questions which probe students' liking for or enjoyment of an activity. For example students may be given positive or negative statements such as:

- I enjoy mathematics (or reading, history, science, etc)
- I don't like mathematics
- I would like to do more mathematics

The students are asked to respond by saying whether they agree strongly, agree, are not sure, disagree or disagree strongly. An example of a series of questions relating to students' social and emotional wellbeing was used by Tan and Bibby (2011) in their research comparing IB and non-IB schools.

An alternative is to ask more open questions such as 'how much do you enjoy mathematics?' to which students respond by choosing a rating from 'very much' to 'not at all'. For younger students this approach can be adapted so that they answer by choosing 'smiley', 'neutral' or 'frowning' faces. Another variation is to ask about particular activities rather than subjects as a whole. Students can be given photographs of other students

involved in various activities and asked how they feel about undertaking such activities themselves.

Approaches using written questions can be adapted for use in interviews with individuals or groups of students. This can provide additional information from body language and allows the meaning of words to be clarified, but some students may be hesitant about revealing their feelings in the presence of others and prefer to give answers in private in writing. Thus several personal factors as well as age have to be considered in selecting the most appropriate methods to use for assessing attitudes in particular cases.

### Observing

In skill and attitude assessment observation can replace questioning as the most appropriate assessment approach. Observation can happen in real time, for example as students discuss a given topic in their lessons, such as 'endangered species', 'global warming' or 'suitable prizes for sports day', or as they carry out an investigation in science, act out a short scene in drama, or measure lengths and weights in mathematics. It might alternatively take place after the event, for example when video recordings of students undertaking practical tasks or engaging in role play are viewed later and performances assessed by teacher raters.

The assessment focus of an observation can range from a small atomistic skill, such as measuring a length or a short time interval, to very complex intellectual problem solving or collaborative social skills (21$^{st}$ century skills). For the latter type of assessment individual or collaborative 'performance tasks' are required that engage learners' interest and that tap the knowledge, skills and behaviours in focus. Teachers observe their students at work on these tasks and informally assess the relevant aspects of learners' development on an ongoing basis, using the evidence to guide further work; over time they build up a range of informal assessments of each student that they can eventually use as a basis for a summative judgement at some future point in time.

When the assessment of performance is summative and the raters are not the students' own teachers, rating rubrics will need to be designed to guide observers as they rate students' performances throughout the task.

Students' attitudes may be inferred from observation of their behaviour when confronted with the objects or situations which are the focus of the attitude of interest. This enables attitudes to be assessed without students being directly asked for their feelings to be expressed or even knowing that the assessment is taking place. For example, as part of the national surveys of science that ran in England, Wales and Northern Ireland in the 1980s, individual 11-year-olds were observed while undertaking three different investigations. As well as recording what the students did, the results they obtained, and the interpretation they made of them, the observers made judgements about their attitude towards being

involved in the investigations, using these criteria for high, medium and low levels of this attitude:

- Shows evidence of real interest in investigation, looking carefully and intently at what happens; actions deliberate and thoughtful.
- Willing to carry out investigation but no sign or great enthusiasm or special interest.
- Carries out only the minimum necessary, may look bored, uninterested or scared. (Harlen et al 1981: 153).

In a sample survey the context of a small number of investigations is too limited and unusual to provide reliable data about individuals and this is not the purpose. However, the general approach of teachers having in mind behavioural indicators of attitudes can enable them to accumulate information about individual students from a number of different activities. These indicators relate to students' willingness to act in certain ways, reflecting their feelings about the objects and situations involved. For example, the suggestions in Box 3.5 are indicators of 'sensitivity towards living things and the environment'.

*Box 3.5 Indicators of attitude towards living things and the environment (adapted from Harlen and Qualter 2009: 187)*

Students who show sensitivity towards living things and the environment:

a) Provide care for living things in the classroom or around the school with minimum supervision.
b) Minimise the impact of their investigations on living things and the environment, by returning objects and organisms studied to their initial conditions.
c) Show care for the local environment by behaviour which protects it from litter, damage and disturbance.
d) Adhere to and/or take part in developing a code of care for the environment, with reasons for the actions identified.
e) Help in ensuring that others know about and observe such a code of care.

Other aspects of attitude may show in other contexts, for instance in the expression of empathy in relation to a character in a story.

### Evaluating products

Product evaluation typically takes place after the event in summative assessment, but in the classroom there could be ongoing formative assessment as the product itself develops. Among numerous possibilities, the product might, for example, be a poster, a three-dimensional model, a wall display, a concept map, a portfolio of work, or a piece of fictional writing. The subject of the assessment might be aspects of writing skill, conceptual understanding, non-verbal communication skills, collaboration skills, aesthetic appreciation, and so on. In order to evaluate a product meaningfully in such an assessment context a set of relevant assessment criteria would need to be agreed, and a rubric designed with which raters could record their judgements. In a summative assessment context it would be

essential to take steps to try to ensure that different raters could apply the criteria in the same way and produce the same rating outcomes for the same product.

## 3.4 The potential for using new technologies in assessment

The value of the use of new technologies in assessment goes beyond assisting in the creation and management of records. Pellegrino et al (2001: 263) point out that it provides 'powerful new tools for meeting many of the challenges inherent in designing and implementing assessment that go beyond conventional practices and tap a broader repertoire of cognitive skills and knowledge.' Many of the current applications of technology in assessment are concerned with the creation, use, scoring and reporting of data from tests, reflecting conventional views of assessment and learning. However, their value in assessment terms is perhaps greater when they are used to support inquiry-based education and reflect a socio-cultural constructivist perspective on learning. The following are some examples of this potential.

- Computer programs can help to increase the effectiveness of formative assessment by helping teachers to analyse students' answers to carefully constructed questions and tasks in greater detail than can effectively be done by the teachers themselves for every student in the class. Programs are also being developed that enable students to test their own understanding, keeping track of students' responses so that the teacher can also monitor the process (Bull et al 2006).

- Concept maps, created by students individually or in groups, are often used to show the relationships that students perceive, and how they understand the links between concepts. When created on paper the teacher has only the product to use in assessing concept development, but computer programs are able to extract more information. Not only can they compare the student's map with a map developed by someone else (more expert) and identify areas needing attention, they can also monitor the communication, collaboration and decision-making as students work to produce their maps (Herl et al 1999; O'Neil and Klein 1997).

- Problem-solving skills can also be explored, though not yet adequately assessed, by programs that analyse the sequence of students' actions as they work though a problem. Realistic and open-ended problems can be presented on screen using graphics, illustrations and sound not easily provided in physical form in the classroom (Griffin et al 2013; Care and Griffin 2011; Chung and Baker 1997).

- Computer-based games are being developed which enable 'stealth assessment' of 21st century competences (for example, system thinking, creative problem-solving, teamwork). Stealth assessment is built into games so that performance data are gathered continuously and used to build up a dynamic model of the learning (Shute 2011).

- Extended written responses can be assessed by programs set up to search for particular words or patterns, which can be used diagnostically – though these are not perfect and their

applicability to the writing of primary-age children must be in question (Magliano and Graesser 2012; Shermis 2010).

# 4. Assessment for learning

*In this section we consider further the nature of formative assessment, including current thinking about how formative assessment should best be implemented in the classroom. Findings from reviews of the academic literature in relation to claims and evidence for the beneficial impact of formative assessment on learning are also presented and discussed. In the final section the role and practice of formative assessment in the PYP are considered in the light of these findings.*

## 4.1 Formative assessment (assessment for learning)

Many of the features of formative assessment (summarised in Box 2.1) have long been implicit in descriptions of effective teaching. Research into students' ideas showed how important it is to take these ideas into account and to involve students in developing their understanding. This led to the assessment of students' ideas and skills in order to inform teaching and to provide feedback to students being recognised as an important role for classroom assessment. Assessment for this purpose of helping learning – as distinct from assessing achievement in order to record and report progress – has been the subject of several empirical investigations since the 1980s. Research studies of classroom assessment have been the subject of several reviews, principally by Natriello (1987), Kulik and Kulik (1987), Crooks (1988), Black (1993) and Black and Wiliam (1998a). The review by Black and Wiliam has attracted attention worldwide and led to further research and to development projects which have explored the implementation of the key components of formative assessment and the impact on students' learning.

The practice of formative assessment, through teachers and students collecting data about learning as it takes place and feeding back information to regulate teaching and learning, is clearly aligned with the goals and practice of inquiry-based learning. It also supports student agency in 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. Thus formative assessment not only promotes learning through inquiry but is a necessary condition for it.

Some of the numerous definitions of formative assessment that have been proposed over the last two decades have been reviewed by Wiliam (2011), who suggests that the main features are brought together in the following definition proposed by Black and Wiliam:

> 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 that was elicited. (Black and Wiliam 2009: 9)

Thus formative assessment is defined in terms of helping to improve the decisions made about how to help learning and, echoing the points about uses made earlier, if the decisions are not improved, then the assessment cannot be described as formative. Nichols et al

(2009) point out a particular strategy or test should not be labelled as being formative unless there are both empirical evidence and reasoned arguments to support the claim of improving student achievement. Some key aspects of formative assessment practice are included in this definition. How these components are combined to support student learning is suggested in the form of a model of formative assessment in figure 4.1. Before looking at the evidence for the impact of formative assessment on students' learning, we consider what the key components are and how they contribute to students' achievement of several important goals of education.

The activities represented by 'A', 'B', and 'C' in figure 4.1 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 and determine what evidence to gather for assessment purposes. For example in a maths lessons a goal might be 'to use a bar chart to represent and record information'. The students' work in activity A might be to create a bar chart about the kinds of pets owned by students in the class. Observation by the teachers and discussion among students in this activity provides opportunity for both teacher and students to obtain evidence of progress towards the goal. However, as well as the content-related goal, the teacher will have in mind other goals concerning how students are learning, and whether they are developing transferable skills. The dialogue, which enables teachers to gain access to students' thinking, should also encourage students to reflect on their learning.

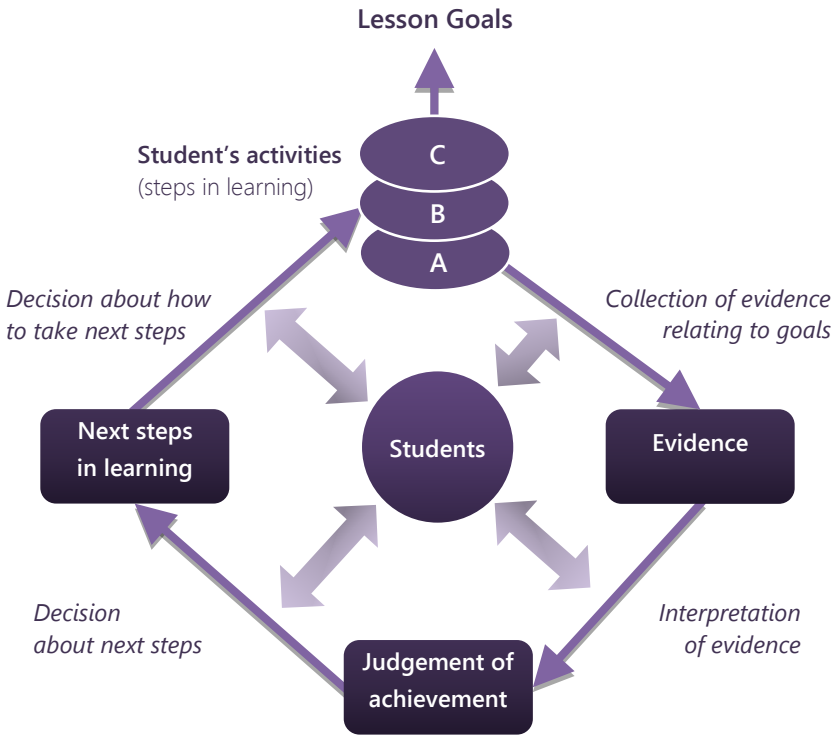


*Figure 4.1 Assessment for formative purposes (adapted from Harlen 2006)*

In order to interpret the evidence from students' work in this example both teacher and students need some understanding of the criteria to apply in assessing the work (can they say what each bar means? does the chart enable them to compare the numbers of certain pets?). Through discussion with students, questioning that elicits their understanding of what they are doing, and listening to how they explain what they are doing, the teacher decides about the relevant next steps, which may be to intervene or simply to move on. As Leahy and Wiliam (2012: 51) point 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'. Activity B is the result of this decision and the source of evidence in a further cycle of eliciting and interpreting evidence.

Students are at 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 in figure 4.1 indicate that students both receive feedback from the teacher and also provide information. They participate in decisions where appropriate through self- and peer-assessment.

The actions indicated by the arrows in figure 4.1 are not 'stages' in a lesson nor necessarily the result of pre-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 might take note of what help is needed and provide it at a later time.

Representing in this way the processes involved that shows formative assessment is integral to teaching. It is part of the practice of teaching that is happening all the time, not something that happens only before, after or at particular points in a lesson.

## 4.2 Implementing formative assessment

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.

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. The aim is to identify – and to help students to take – next steps in progress of developing their concepts, knowledge, skills and attitudes.

The role of the teacher in formative assessment is to put into practice the key features of formative assessment as indicated in Box 2.1 in section 2. Essentially these are:

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

The following brief points about each of these features reflect current views on how to implement them in a manner consistent with inquiry-based learning and the view of learning on which it is based. [For a more detailed and extensive description of formative assessment strategies, see Wiliam 2011.]

### Questioning

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 questions that are likely to give teachers access to students' thinking are 'open' questions, which encourage students to express their ideas rather than to answer a point raised by the teacher, and 'person-centred' questions which directly ask for the students' ideas and conjectures ('what *do you think* is the reason for…') rather than asking for facts ('what is the reason for …'). These are also the type of questions that stimulate students' thinking. In relation to *content*, questions need to be matched to the purpose of asking. There should be a reason for the question and interest in the answer. If the answer is to be useful in developing thinking then it has to give the kind of information or stimulate the kind of response required. For instance, a question to encourage close observation might be 'what do you notice that is the same about …?'. To encourage collaboration 'how are you going to decide which of the ideas your group can suggest is the one that works best?'. Teachers might think of some questions and how they should be worded as part of their lesson planning.

Carefully worded questions designed to stimulate thinking deserve thoughtful answers and students need to be allowed time to respond to such questions. 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 using formative assessment to help learning through inquiry. Research into questioning has provided some ways in which teachers can signal to students that a thoughtful answer is expected (Carlsen 1991). One strategy is to increase the 'wait time', the time between asking a question and receiving a response. Teachers often expect an answer too quickly and in doing so deter students from thinking. Research by Budd-Rowe (1974) showed that extending the time that a teacher waits for students to answer has a

marked effect on the quality of the answers, a finding which has been confirmed by more recent research (Black et al 2003; Tincani and Crozier 2007). Other advice is to avoid rephrasing a question if it is not answered straight away and instead asking students to discuss the question with a partner or group before calling on individuals to give an answer.

### Classroom dialogue

There is abundant research in support of the view expressed by Alexander (2008: 10) that 'Language and thought are intimately connected, and the extent and manner of children's cognitive development depend to a considerable degree on the forms and contexts of language which they have encountered and used'. However, as recognised in a socio-cultural constructivist view of learning, meaning is constructed by learners through interaction with others. It is in trying to find words to communicate their meaning to others that students have to reformulate their ideas in ways that are influenced by the meaning that others give to words. The teacher's role is to encourage this interaction by taking part in the dialogue in a way that encourages students to clarify their meaning, base arguments on evidence and take their thinking to a deeper level. Alexander has described this role as 'dialogic teaching', through which teachers can 'steer classroom talk with specific educational goals in mind'. He cites neuroscience as supporting this active role of the teacher in two ways:

> … first in confirming the importance of teaching as intervention rather than mere facilitation; secondly as an endorsement of teaching which capitalises on the collective and interactive environment which classrooms offer. (Alexander 2008: 13)

In inquiry one of the ways in which teachers 'steer' the dialogue focuses on the use of evidence and may lead to what has been described as 'argumentation'. How this differs from argument in daily life is explained in the context of primary science by Michaels et al (2008):

> 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. (Michaels et al 2008: 89)

### Feedback to students

Feedback has a key role in formative assessment since it is the mechanism by which future learning opportunities are affected by previous learning. It has been described by Hattie and Timperley (2007) as 'one of the most powerful influences on learning and achievement', but they warn that whether or not it has a positive effect on learning depends on several factors since the process is complex. Feedback is most obviously given by teachers to students orally or in writing, but also, perhaps unconsciously, by gesture, intonation and indeed by action, such as when assigning tasks to students. Teachers' views of learning influence the form in which they provide feedback and decide its content. Constructivist views of learning lead to interaction between teacher

and students in which students respond to the teacher's comments and suggestions rather than the one-sided communication from teacher to student that is typical of a behaviourist view of learning.

Current views of the form that feedback should take have been strongly influenced by the research of Butler (1988). In a well-designed and complex study, which involved different kinds of task and students of different levels of achievement, Butler compared the traditional feedback on students' work in the form of a grade or mark with two other forms, one giving comments on the work and how to improve it but no mark, and the other giving such comments and a mark. The results showed that feedback in terms of comments alone 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. 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.

Of course the nature of the feedback is important, since, as Wiliam (2010: 144) points out from a review of research, 'Just giving students feedback about current achievements produces relatively little benefit, but when feedback engages students in mindful activity, the effects on learning can be profound'. The main features of effective feedback were giving students explanations and specific activities to undertake to improve. Although many of the studies of feedback have concerned older students Wiliam found that 'attitudes to learning are shaped by the feedback that they receive from a very early age' (ibid). The evidence from research and practice (e.g. Tunstall and Gipps 1996a, 1996b) indicates an important difference between feedback that gives information and feedback that is judgemental. This applies to feedback given orally as well as in writing. Feedback giving 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.

In contrast, feedback that is judgemental is expressed in terms of how well the *student* has done (this includes praise as well as criticism) rather than how well the *work* has been done, and gives a judgement that encourages students to label themselves and compare themselves with others.

In summary, research points to the following conditions necessary if feedback to students is to help learning:

- 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.

However, Black et al (2003) found a considerable gap between the classroom practices necessary for students to benefit from teachers' comments and what actually happens.

### Using feedback to regulate teaching

In formative assessment the feedback that teachers receive from observing their students' response behaviours is used in making decisions about how to help them take their next steps in learning. This feedback enables teaching to be regulated so that the pace of moving towards the learning goals is adjusted to ensure the students' active participation. Regulation in this context, as in all regulated processes, ensures effective operation by constantly monitoring the effect of change and responding to it. Since 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. An important source of feedback to the teacher comes from students' self-assessment and peer-assessment, since the criteria they use in judging the success of their work reflects their understanding of what they are trying to do.

### Encouraging student self and peer assessment of their work

Learning through inquiry and formative assessment share a common aim of students becoming increasingly able to take part in decisions about their work and how to judge its quality. Students 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 4.1. In order to assess their work students need to know its goal or aim and the criteria by which to judge its quality. To communicate to students the goal of an activity means giving a reason for it (usually in the form of tackling a problem or question), not telling students what to do or what they should learn. 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 what they are trying to achieve, can be reinforced by dialogue and questions during the activity and in later discussion of what was done and found.

Communicating a view of 'quality', or the standard of work to aim for, is part of the feedback which teachers give to students. It can also be made more explicit by discussing with students anonymous examples of work and identifying the features that make some more successful than others. Another approach is to brainstorm with students the criteria that work of certain kinds (such as reporting a science investigation, writing a story, presenting an argument) should meet. Students then use the agreed list as guidance and as quality assessment criteria when relevant work is being undertaken. This illustrates one way in which students can be helped to consider the nature of their learning and reflect metacognitively on how to improve it.

Peer assessment, i.e. involving students in assessing each other's work, as part of formative assessment, has been strongly advocated by Sadler (1989). Black et al (2003) provide several arguments in favour of encouraging students to judge each other's 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. The feedback that students give to each other can also be more effective since it is 'in language that students themselves would naturally use' (p50). The interchange between students also helps them to understand what they are trying to do and what success involves (Stobart 2008: 184).

A caveat has to be noted, however, relating to the classroom ethos and power relationships within some groups of students. Case studies by Crossouard (2012) of peer assessment by students aged 11 and 12 provides disturbing evidence of how gender, social class and student attainment hierarchies are 'implicated in processes that are typically bathed in the supposed 'neutrality' of assessment judgements.' (p736). 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 need to recognise the issues of equity that may be raised when practising peer assessment.

## 4.3 Impact on learning

The importance of formative assessment lies in the evidence of its effectiveness in improving learning and in arguments that it leads to competences that are needed for continued learning. Empirical studies of classroom assessment have been the subject of several research reviews. As mentioned earlier, the review by Black and Wiliam (1998a) attracted particular attention. This was partly because of the attempt to quantify the impact of using formative assessment. In a summary of their review findings, the authors estimated that the quantitative studies included in their review produced 'effect sizes' of between 0.4 and 0.7 and noted that 'such effect sizes are larger than most of those found in educational interventions.' (Black and Wiliam 1998b: 4). Further, they reported that 'improved formative assessment helps the (so-called) low attaining students more than the rest, and so reduces the spread of attainment whilst also raising it overall' (ibid).

Black et al (2003) cite research by Bergan et al (1991), White and Frederiksen (1998) and the review of Fuchs and Fuchs (1986) as providing evidence of better learning when teachers take care to review information about students and to use it to guide their teaching. Butler (1988) showed the importance of non-judgemental feedback in the form of comments with no marks. Schunk (1996) also found positive impacts on achievement as a result of students' self-assessment. These all reflect a view of learning in which students participate actively rather than being passive receivers of knowledge. This means that assessment used to help learning plays a particularly important part in the achievement of the kinds of goals of understanding and thinking valued in education for the 21st century.

A number of other more recent reviews and investigations (e.g. Brookhart 2007; Hattie and Timperley 2007; Shute 2008; Wiliam 2009) have added to the evidence of positive impact, leading Leahy and Wiliam to claim that:

> 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%. (Leahy and Wiliam 2012: 52)

Stobart (2008: 154), however, strikes a note of caution, pointing out 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. It can, of course, 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 rather slender base of evidence was also noted by Briggs et al (2012). As part of their critique of a meta-analysis by Kingston and Nash (2011) of studies of the effect on learning, they concluded that 'the hype and marketing of formative assessment has greatly outstripped the empirical research base that should be used to guide its implementation' (p16). Shepard (2009: 36) also warns that the powers of formative assessment to 'raise student achievement have been touted, however, without attention to the research on which these claims were based'. In addition, Bennett (2011) has criticised the loose conceptualisation of formative assessment.

However, as well as evidence of effectiveness, there are compelling theoretical arguments for the importance of formative assessment, based on the widely accepted theories of learning that emphasise the role of learners in constructing their own understanding. 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 (Bransford et al 2000). The feedback to teachers provided through formative assessment has a role in regulating teaching so that the pace of moving forward is adjusted to ensure the active participation of the learners. Feedback to students in formative assessment enables them to recognise where they are in progress towards goals and participate in decisions about their next steps in learning.

## 4.4 The PYP approach to formative assessment

The PYP perspective on assessment places considerable emphasis on formative assessment, or assessment for learning. The review of current thinking about assessment in the academic literature and a growing body of research gives full support to this emphasis. However, formative assessment is not a single or simple tool but a collection of strategies generally taken to include teachers using particular forms of questioning, engaging students

in discussing their ideas and basing their arguments on evidence, providing feedback to students, using feedback to adjust teaching, and involving students in assessment of their work.

Whilst the various aspects of formative assessment have support from current views of learning – and in particular as to how the skills, knowledge, conceptual understanding and attitudes of 21st century education can be best achieved – the research evidence comes from studies of the impact of changing separate aspects rather than the combination of changes in questioning, dialogue, feedback and student self-assessment. Nevertheless, it is important to know that change in each of these is effective in improving student achievement since it is hard for teachers to change several aspects of their practice at the same time (see Gardner et al 2010 for accounts of implementing change in assessment). What happens in practice (Black et al 2003) is that teachers choose to focus on one aspect (often their questioning) and then, motivated by the response of their students, extend the changes to other aspects.

Given the challenges facing many teachers in making changes in their teaching as they plan to implement formative assessment, the PYP might consider providing some focused professional development in this matter. For example, once introduced to the range of practices involved, teachers might be offered a simple self-reflection checklist to identify where they could most effectively begin to make change, perhaps in the nature of the feedback they provide to students or discussing quality criteria that students could apply to their own work.

In relation to the overall conception of formative assessment, it needs to be made clear that this use of assessment is integral to the way of teaching rather than an activity which can stand alone and be linked, or not, to teaching. For instance the statement on page 45 of *Making the PYP Happen* that 'Formative assessment and teaching are directly linked and function purposefully together' suggests a potential separateness that is probably not intended. Using formative assessment means that teachers do what they have to do anyway in a way that helps learning: all teachers ask questions but it is the kind of question that makes the assessment formative; all teachers give feedback to students but it is the content and form that determines whether it will help the students' learning. Research studies, supported by arguments about how learning takes place, have identified the particular aspects of teacher-student interaction that enable assessment to be used formatively.

In *Making the PYP Happen* the prime objective of assessment is stated as being 'to provide feedback on the learning process.' Feedback is a key feature of formative assessment which has been particularly well researched. This applies mainly to feedback to students, where, as noted earlier, the evidence has highlighted the importance of giving this in the form of comments rather than marks or grades. At the same time it is important that the comments provide guidance for improvement or moving on, that they are non-judgemental, focused on the work rather than the student, and that students are given time to respond to the

feedback. Feedback of this kind is the essence of using assessment to help learning and it is appropriate that the PYP assessment gives particular attention to it. However, the accounts in the literature of a gap between classroom practice and what is needed for students to benefit from good feedback suggests that it may be necessary to spell out more explicitly and with exemplification just what is involved in providing and using it effectively. In this regard it is worth repeating the conditions for using feedback to students to help learning, suggested by research:

- 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.

However, in formative assessment, information gathered about students' learning is also fed back into teaching, being used to adjust – or regulate – teaching so that the demands made on students are neither too severe nor too simple. Adjustment may be necessary in order to maintain students' active engagement in learning which is less likely if they are bored with their tasks or feel unable to understand what is required. Teachers need to be constantly monitoring their students' reactions in order to decide the next steps both for the students and for themselves.

For example, Harlen and Qualter ( 2009) cite a teacher of 6 and 7 year olds who transformed part of her classroom into an 'iceberg' as a context for investigating conditions in which ice can be kept from melting. She gave them a simple worksheet for planning, and recording, but as she observed their activities and talked to them she realised that many were not yet clear enough about how their ideas could be investigated to be able to express them in writing. So she decided in the next lesson to focus on discussing and clarifying what was involved in carrying out tests of the ideas they had about keeping the ice from melting. Rather than letting the children struggle, she decided to change her plans.

Although teachers' experience and careful planning will avoid any major mismatch between students' learning tasks and their ability to engage with them, if feedback from students is taken seriously it will not always be possible to follow the pre-planned course of a lesson and some flexibility will be needed.

The PYP documentation on planning for assessment notes the importance of including ways of finding out students' ideas about the particular context and focus of the inquiry. This enables teachers to gather the information that enables them to help students towards achieving the particular goals of the lesson. The need to focus this data gathering and to 'constantly look for evidence that meets the criteria' is noted, but more emphasis ought to be given to how this information will be used to help learning. Finding out about students'

understanding, knowledge and skills fulfils the need of a constructivist approach in teaching, but it is not the main aim in formative assessment; rather this is to use this information to decide the next steps for the students and how best to help them take these steps.

The process of deciding next steps can be helped by an aspect of formative assessment practice that deserves more attention in the PYP planning for assessment. Although the PYP Practices relating to Standard C4 include 'opportunities for students to participate in, and reflect on, the assessment of their work' (*Programme Standards and Practices*, 2010) there is little guidance as to what 'participation' might mean for students at different stages and ages. Students take part in assessment anyway, of course, by answering teachers' questions and sharing in various ways the ideas they have, but more deliberate and conscious involvement has several advantages for their learning. It not only adds to the information about their understanding that teachers can use in helping learning but it also enables students to take some responsibility for it. Since it is the students who have to take the next steps towards the goals of their work, their participation in deciding these next steps makes it more likely that they understand what they have to do and will be committed to putting in the necessary effort. For this to happen students have to be aware of the goals of their work and the criteria of quality to be applied in assessing it. Ensuring this understanding is entirely consistent with the PYP philosophy and so we suggest that this aspect of formative assessment practice is given more attention.

# 5. Assessment of learning

*In this section we consider further the nature of summative assessment. The principal ways in which summative assessments are arrived at by teachers are overviewed, including strategies for eliciting evidence of achievement from special tests and tasks, using portfolios of students' work and summarising teachers' observations. The issues of assessment validity and reliability are also addressed. Against this background we discuss the PYP approach to summative assessment for recording and reporting students' achievement and the role of the exhibition.*

## 5.1 The nature of summative assessment

In this section the concern is with assessment of individual students' conceptual understanding, knowledge, skills, attitudes and action for the purpose of recording and reporting development at a particular time. This covers a vast range of topics and in order to keep to manageable proportions, the focus is on the outcomes of primary school students' learning through inquiry. This focus does not mean that basic knowledge of facts, vocabulary and conventions are not assessed. But there are many familiar ways of doing this through classroom tasks, tests and quizzes devised or chosen by the teacher and used at appropriate times. More challenging is to ensure that conceptual understanding, inquiry skills, reasoning and use of evidence, together with associated attitudes and action, are assessed. If these attributes are neglected in summative assessment then, on account of the relationships indicated in figure 5.1, it is likely that they will be neglected in the content and pedagogy of the curriculum experienced by students. Martone and Sireci (2009) discuss the importance of alignment among these aspects of the curriculum and describe methods for evaluating the alignment between state standards and assessment.

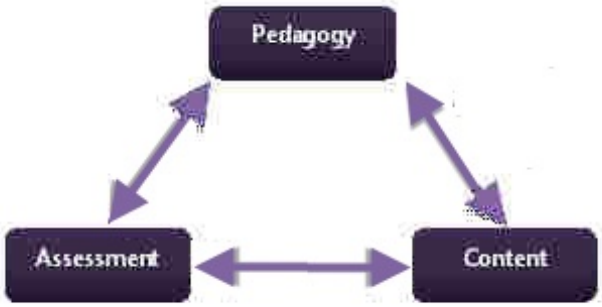


*Figure 5.1 Interactions between assessment, pedagogy and curriculum content (from Harlen 2013)*

Figure 5.2 represents the process of summative assessment that is mainly carried out for the purpose of reporting achievement. There is no direct feedback into learning *as it takes place* as there is in formative assessment, indicated by the closed cycle in figure 4.1 in section 4. However, the information about achievement may be used in various ways which have some formative function: it may be used to inform decisions about the future learning opportunities of individual learners; it may be fed back into decisions about teaching and

about the programme of study; and it may contribute to educational policy making. These are the three levels of use that Stiggins (2001: 31) identifies for assessment results, the first level being use in the classroom, the second being use for 'instructional support', and the third being use by policy makers. Use at the classroom level includes summative as well as formative assessment, and Black et al (2003) suggest ways in which what are essentially summative tests can be used formatively. This may well help learning, through future planning, but not in the direct way that is the purpose of formative assessment. Moreover, the authors warn that 'the pressures exerted by external testing and assessment requirements are not fully consistent with good formative practices.' (Black et al 2003: 56)

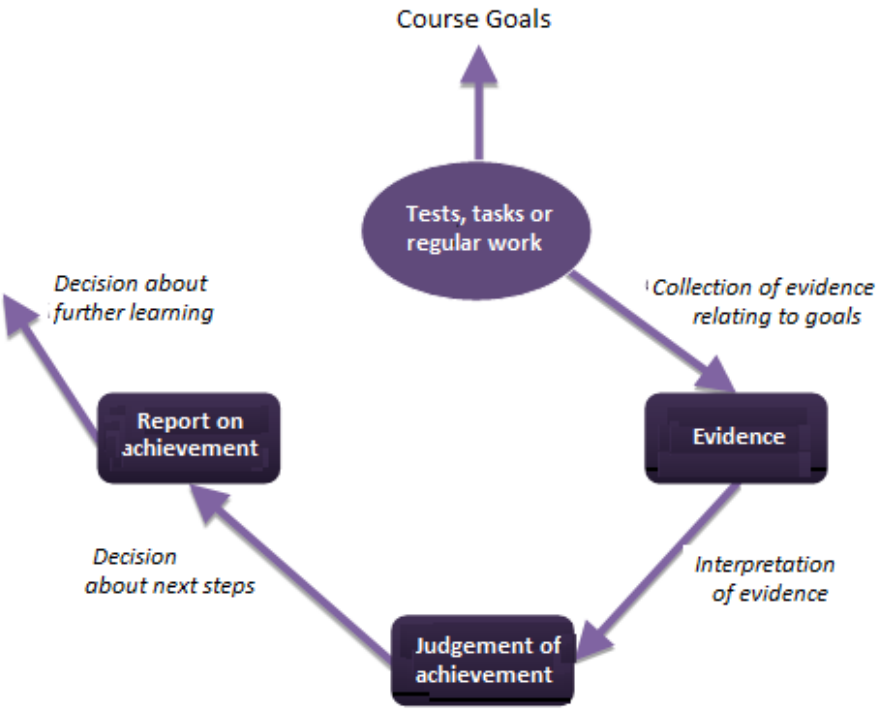


*Figure 5.2 Assessment for summative purposes (adapted from Harlen 2006)*

As noted in section 3, in any assessment context evidence is collected from tests, tasks, special activities or from regular work, designed to enable the collection of data relating to the overall goals of a course or extended unit of work. The data can be collected by a range of means from different sources: students' writing or drawing, artefacts constructed by students, portfolios, observation of actions, discussion or presentations of work. Interpretation is 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. The 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. Students are all judged by the same criteria, embodied in mark schemes (rubrics), whereas in formative assessment judgements are often made in terms of the progress made by individual students from different starting points.

The process of interpretation inevitably replaces the richness of the actual performance by 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 optimising reliability within any given constraints. Where results are used to compare students, particularly where high stakes selection or grading is involved, steps are taken to check marking and to moderate judgements by teachers or examiners. The concern for reliability in high stakes testing can often lead to reduced construct validity, as noted in section 2, due to the unavoidable trade-off between reliability and validity in a context of resource limitation and assessment manageability. However, 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 achievement of a wider range of skills, knowledge and competences.

The form of reporting depends to a large extent on the nature of the task, the basis for judgement and the audience for the report. In general, a combination of assessment strategies will be needed to reflect the development of the whole child. In the case of tests, numerical scores are a summation over a diverse set of questions. The same total can be achieved in many ways, so scores have limited meaning for what students actually know or can do. They also give a spurious impression of precision, which is 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.

### Assessing concepts and skills within inquiry-based learning

When attempting to assess understanding and inquiry skills any task used should:

- involve students in working on some particular aspects of inquiry;
- be new to students so that they are using their knowledge or skill and not simply recall of information, reasons or algorithms that have been committed to memory;
- interest and engage the students so that they want to do their best to tackle it.

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 skills, such as the transdisciplinary skills identified in the PYP (thinking skills, social skills, communication skills, and research skills such as observing, collecting and interpreting data). However, the assessment of conceptual understanding and skills cannot be entirely separated from each other. Since skills have to be used in relation to some subject matter, some knowledge of the subject matter will be involved. Similarly, tasks used for assessing understanding will require some use of skills (such as reading, explaining, interpreting, presenting). Thus there will always be some aspects of understanding and of skill required in *all* tasks. What determines whether a task is primarily assessing conceptual understanding or skill will be the level of demand on one or the other, and the credit represented by a score given to different kinds of response.

With regard to summative assessment, information about students' performance at a particular time can be obtained by

- tests or special tasks given under controlled conditions or embedded in classroom activities;
- building a record over time, as in a portfolio created by teachers and/or their students;
- summarising information gathered by teachers during their work with the students over a period of time.

We now consider each of these in relation to their suitability for summative assessment of conceptual understanding and skills for students of primary school age.

## 5.2 Using tests or special tasks

The use of tests to find out what a student knows or can do is a time-honoured approach to summative assessment. It is an attractive approach for certain purposes, particularly where students are put in competitive environments, because each can be given the same task or asked the same questions in the same conditions. This contrasts with using teachers' judgements of regular work where the evidence arises from situations which might vary from student to student. However, it should be noted that giving the same task does not necessarily provide the same opportunities for students to show what they can do.

There is a wide range of types of test, with various kinds of task (written or performance) and ways of responding, as noted in section 3. In the case of written questions, questions expressed in an open form are better suited to assessing conceptual understanding by eliciting what students know and understand than questions where they choose between given answers. Multiple-choice items are open to guessing and clues given by the words used in the options. However, open response questions also have disadvantages, particularly for younger children who have greater difficulty in being as specific in writing as they might be in talking. Interpreting students' written answers can also be a problem. Even when these are not the focus of the assessment, there will be some dependence on the students' reading and writing skills, especially where an attempt is made to set the question in a context that will seem interesting and relevant to the students. In addition, younger learners can be quite vulnerable to test fatigue, and the requirement to write an answer in more than a word or two can deter them from attempting an answer at all.

In order to reduce the burden on reading and writing skills, where these are not the focus of assessment, students can be observed whilst performing a task in a real practical context. Skills relating to performance in music, drama, oral presentations, and investigations in science and mathematics can be assessed in this way, and often need to be in the interest of assessment validity, with students working either individually or in groups.

Performance tasks, although having high face validity, have some disadvantages in relation to reliability. They are generally extended tasks, taking time to administer, thus few such tasks can be given to any individual student. This means that the particular content of the task might favour some students, who engage readily with the chosen content, but

disadvantage others who find it less engaging. The same issue arises in written questions and tasks since there are many different contexts that could be used to assess a particular skill or concept and those used in a particular test are only a sample of the possibilities. If a different sample of questions is chosen, some students would find some of them easier to answer but others might find them more difficult. So a different selection would produce a different result for different students, giving rise to what is described as 'sampling error' (a contributor to unreliability), which is typically greater the fewer items there are in the test. As a result there is considerable uncertainty about what might be the 'correct' result for individual students. Indeed, Black and Wiliam (2012: 254) estimate that in the case of a national reading test for 11 year olds, the score of a student given 64% might vary between 52% and 76%.

## 5.3 A portfolio built over time

This form of portfolio is not a sample of *all* a student's work over a period of time, but reflects the best performance at the time of reporting. The evidence is accumulated gradually by retaining what is best at any time in a folder, or other form of portfolio (including computer files), and replacing pieces with better evidence as it is produced. Whilst it is important to base the summative judgement on evidence from a range of activities and not judge from one task, there is no point in including work that no longer reflects what students are capable of producing. Such an approach enables students to have a role in their summative assessment by taking part in the selection of items in the folder or portfolio, a process for which they need some understanding of the broad goals and quality criteria by which their work will be judged. It is important that time is set aside at regular intervals specifically for students to review their work. This gives them time not only to decide what to put in the 'best work portfolio' but also to consider what they can improve.

The number of items in the portfolio grows from the start, but imposing a limit (of around 5 to 10 pieces for each subject or unit of inquiry) ensures that care is taken in decisions and that students really think about how they are judging their work. In this way formative assessment is at the centre of the approach. When it comes to the end of the year or period of reporting, the teacher has the information in the portfolio to write a narrative report on each student's performance in each area of work for parents and for the student's next teacher, and also to summarise achievement of unit and course goals as necessary for school records and programme evaluation.

Summary judgements made by teachers in relation to whether students have achieved overall goals need to be moderated, ideally by groups of teachers who meet to review sample portfolios and judgements made on them. Where this is practised it has value not only in potentially assuring greater reliability of the assessment but also in providing professional development (e.g. Maxwell 2004).

The use of portfolios is often criticised for the accumulation of large amounts of material, but already there are examples of avoiding this by creating portfolios electronically. The

rapid spread of the use in primary classrooms of mini-laptop computers linked wirelessly to the teacher's computer, opens up the potential for assessment to be carried out without the need for a physical portfolio. The process has been developed and trialled by Kimbell et al (2009) in the E-SCAPE (Electronic Solutions for Creative Assessment in Portfolio Environments) project, initially for the assessment of technology projects, but later extended to other subjects. Students send their work, in the form of text, voice or video recordings, photographs or drawings, to the teacher as the work progresses as well as at the end of an activity. The uploaded work can be discussed with the student electronically or in person and the collection could be edited or changed as it is superseded by later work. The teacher stores the collection and uses it for end-of-year reporting as in the case of a physical folder. Moderation of teachers' judgements for school records can be carried out using the electronic form.

## 5.4 Summarising teacher-based assessment

Since teachers are assessing students' work through the year, the resulting evidence of developing understanding and skills can be brought together at the end of the year, or other reporting period, to summarise their achievement. Ideally the information will have been gathered in the context of formative assessment and used to help learning. However, it is not appropriate to rely on the interpretations made for formative purposes, and to transfer them directly to summative assessment, since these will have taken into account factors relating to individual students' progress and effort. Whilst this does not matter for formative assessment, where identifying the levels at which students are working is not necessary, in summative assessment what has been achieved has to be judged against the same criteria for all students.

Whilst the *evidence* used for the two purposes might be the same, the judgements in summative assessment are made differently. Where there are 'levels' specified, the evidence has to be reviewed against the broader criteria that define the levels. This is usually done by the 'best fit' method, recognising that there will never be a perfect match between the evidence and the criteria used in judging it (see Johnson 2013 for comment on the reliability of such summative teacher assessment in high-stakes applications). An example of this approach is found in the reporting of the achievement of students aged 3 -5 in England, shown in Box 5.1.

In making these decisions about students' achievements at the end of the Reception Year, just prior to entering formal primary education, practitioners in England are expected to refer to the exemplification material published on its website by the Department for Education. This material provides examples of observation notes made by practitioners, students' quoted words, students' pictures, photographs and in some case comments of parents, all of which illustrate a student's learning and development which best fits the 'expected' category.

*Box 5.1 Summarising student achievement in England in the Early Years Foundation Stage Profile (STA 2012)*

The Early Years Foundation Stage (EYFS) Profile is a summative record of children's attainment, completed by Early Years practitioners in the final term of the year in which the child reaches the age of five.

The profile records practitioners' judgements in relation to three primary areas of learning (communication and language; physical development; and personal, social and emotional development) and four specific areas (literacy, mathematics, understanding the world, and expressive arts and design). The primary and specific areas of learning are sub-divided into a total of 17 statutory Early Learning Goals each with an explanatory note. For example, for 'understanding the world' the Early Learning Goal is:

"Children know about similarities and differences in relation to places, objects, materials and living things. They talk about the features of their own immediate environment and how environments might vary from one to another. They make observations of animals and plants and explain why some things occur, and talk about changes."

Practitioners use their observations and records made of each student during the EYFS to judge the level of development in each of the 17 Early Learning Goals as:

- 'emerging' (not yet at the level of development expected at the end of the EYFS)
- or 'expected' (meeting the description of the level of development expected at the end of the EYFS)
- or 'exceeding' (beyond the level of development expected at the end of the EYFS).

In theory, reporting against criteria which describe performance at progressive levels or grades can provide a more meaningful indication of what students have achieved than a test score. Rather than a single overall grade or level, which would have to combine different domains, a profile is preferable as in reporting language achievements separately for listening/speaking, viewing/presenting, reading and writing. The shorthand of 'levels' – numerical or alphabetical labels given to progressive criteria – can be useful for some purposes, but when reporting to parents and students it is necessary that the meaning of the levels is explained and preferably accompanied by accounts of what the student can do. Moreover it has been recognised that the use of levels can have negative implications for students' motivation and learning. For example, after 25 years of structuring the curriculum and reporting in terms of levels in England, the use of levels has been challenged in the course of recent curriculum and assessment revision. The reasons, which no doubt resonate in other systems, were set out by an influential expert group in 2011 and summarised in Box 5.2.

*Box 5.2 Some negative impacts of reporting achievement in terms of levels (extracted from DfE 2011)*

- 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 (high stakes), 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.
- As level are generally well spaced (usually 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.

It appears that, whether assessed by test or by teacher judgement, reporting students' performance as a 'level' of achievement has been found to have a profound impact on how teachers, parents and students themselves judge their progress and themselves. It underlines the point that any written records are more meaningful if accompanied by face-to-face communication through conferences. Such conferences also allow interaction that facilitates reporting of a range of achievements including progress in the attributes of the learner profile.

The absence of reference to students in figure 5.2 implies that that they do not have a role in summative assessment, and generally this is indeed the case. Ideally, however, the formative feedback that students receive during the course of their work enables them to realise what they have achieved and what they need to do to make further progress. When the process of summative assessment is an open one, not restricted to what can be done in a test or controlled situation, and assessment criteria are shared with students and users of the results, there is a greater 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, but there should be no surprises when their performance is summarised.

One of the reasons for testing students is that the results can be used at national or regional level to monitor the achievement of groups of students' and in many cases, set targets for

schools, with penalties when these are not met. This 'high stakes' use of test results puts pressure on teachers (ARG 2002b; Harlen and Deakin Crick 2003), 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. These findings also raise questions about equity, since negative responses to tests are not spread evenly across all students. Some 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 high-stakes national tests students are subjected to 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. (Alexander 2010: 316)

Moreover, Torrance (2011) points out that even though national results have improved, much evidence suggests that, if anything, actual standards of achievement are falling, and grade inflation is undermining public confidence in the whole system.

Similar effects are reported from other jurisdictions where the results of tests are used to make judgements on schools and teachers. (OECD 2013)

## 5.5 The PYP approach to summative assessment

Summative assessment can be used in several ways, from simply reporting on an individual student's achievements at the end of a course, to filtering students into different future educational pathways, to evaluating the effectiveness of the curriculum, or of teachers, schools or entire educational systems. The discussion here focuses on reporting to parents and keeping records that enable students' progress in learning across a range of goals to be monitored.

The form the individual assessment result takes may be appropriate for one use but less so for another. For example, what is most useful to parents is a rich description of what their child has achieved, expressed in narrative terms, related to the curriculum and indicating further lines of development. Some schools provide reports for parents that include a section written by the students, saying what they have enjoyed, feel confident about and what is less secure. Ideally this written report is supplemented by face-to-face discussions involving parent, student and teachers, as indicated in the PYP reporting guidelines (IB 2007: 51).

Such rich reporting is more appropriate for parents than a list of marks, grades and labels which are not immediately meaningful in terms of what the student knows, understands and can do. However, narrative descriptions that communicate well with parents are not easily summarised for the purpose of school records and use in curriculum evaluation. For that purpose a student's achievement in particular curriculum areas has to be reduced to a judgement of the extent to which expected learning outcomes have been achieved. This need not be in quantitative terms, but could be a simple judgement of learning outcomes 'not yet achieved', 'achieved' or 'exceeded'. Thus the form as well as the content of the summative assessment has to match the intended use.

In whatever form the judgement of achievement is expressed, it has to meet certain standards of reliability and validity appropriate to its use. Validity is paramount in any assessment exercise, and however teachers record the summative judgements of their students they should be applying the same criteria in the same way to arrive at those judgements.

For practical reasons the issue of reliability is essentially irrelevant where narrative descriptions of an individual student's achievements are produced by a class teacher for the purpose of reporting to parents. This is because it is not possible to know whether an alternative account might have been made had the student been taught by a different teacher throughout that term or year. However, if the summative assessment for the student is recorded in terms of marks or grades, and if the marks or grades achieved are used to make decisions about educational choices further up the school system, then the dependability of those assessment results becomes a very important issue. The degree to which an understanding of assessment criteria is shared by teachers, and the extent to which standards of judgement are aligned, ought to be checked empirically from time to time using some form of moderation.

The PYP unit planning guide for teachers includes a section for teachers to set out 'possible ways of assessing student learning in the context of the lines of inquiry' and to identify the evidence to be sought. The outcomes of the unit are likely to be varied in form – including individual and group written work and artefacts such as charts, posters, photographs, video-recording, and actions. The process of judging students' achievement involves deciding how well information about their performance meets the criteria used to assess it. There are three important aspects to consider here: the collection of information about student learning, the criteria to be applied, and how to bring together the information and criteria in making the judgement. The assessment is carried out by teachers, and whilst there is guidance in *Making the PYP Happen* about choosing tools and strategies to gather information, and the annotated samples on the OCC provide some examples of criteria, there is little to help teachers create criteria for specific outcomes or to apply the criteria to the information, for example, by using a 'best fit' approach (see section 5.4).

The sample assessment documents serve a variety of purposes. Some relate to formative assessment, some to reporting, some to recording (keeping a class record of judgements about progress in various skills) and some to student self-assessment. Those which do relate to summative assessment of a unit show a plethora of different approaches to identifying assessment criteria. For example, the rubric for a grade 5 mathematics unit provides detailed criteria for each specific goal of the unit expressed at four levels: 'beginning', 'consolidating', 'proficient' and 'advanced'. In the sample from another school, also a grade 5 unit, the rubric used set out criteria at three levels, simply described as 1, 2 and 3. Of course there is no way in which a 'level 2' for a grade 5 student in the second school bears any relation to one of the four levels at the other school. In the mathematics class, the graded criteria were shared with the students, thus giving the rubric a formative role. The issues arising from individual schools identifying assessment criteria without a common framework also arise in relation to the exhibition and are further explored in that context below.

Notably absent from the samples is explicit reference to the assessment of action. It could be argued that assessment of action by students as a result of their inquiries in the unit or in the exhibition is not required; the important point being that some action is taken. However, this should not mean that 'anything goes' and that the nature and relevance of the action is unimportant. Indeed it is suggested in *Towards a Continuum of International Education* that sometimes 'inaction is the best choice'. The main point is that a conscious choice has been made by the student, based on reflection of the work in the unit. It may be enough to spell out criteria for judging whether the action (or inaction) has added further to the evidence of learning from the unit, but further consideration should be given to providing guidance on how relevance and adequacy of the action is judged if this important aspect of the PYP is to be taken seriously.

Some experience of procedures that have been found to help teachers in developing their assessment practices are reported in research in Australia by Skamp (2012). The *Primary Connections Project* (Australian Academy of Science), developed between 2004 and 2012, aims at developing primary school 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 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. In the units 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 that learning can be monitored. Strategies for peer- and self-assessment are also included.

The relevant findings come from the report on the 'Evaluate' phase of the lessons, whose purposes 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.

Skamp's report of feedback from teachers found that 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, suggesting that this aspect of the lessons may not have been implemented by many teachers. Those who did respond 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.

Overall, while it was clear that teacher-based assessment was not strong, the research showed some teachers gradually building their confidence in this aspect of their work. 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.

This suggests that both teachers and students may need help in identifying students' use of, and progression in, inquiry skills. It also appears to be helpful to make explicit the view of learning that underpins the assessment processes and to provide short examples online or on CDs of the processes in action.

## 5.6 The Exhibition

The PYP exhibition enables students to bring together what they have learned from the inquiry units and other elements of the curriculum. In this way the experience of the PYP helps teachers consciously to build students' understanding into powerful ideas and essential skills and attitudes, thereby ensuring that the students arrive at a picture of the world that is not a collection of independent assertions but parts that connect with each other. It avoids learning becoming fragmented, recognising that just as a house is not a pile of bricks so learning is not a pile of disconnected facts. Thus the inclusion of experiences, such as the exhibition provides, finds considerable support in views of modern education.

Two features in particular reflect current understanding of what is important in learning and should be retained in any revision of the exhibition. The first is the insistence on group work, which ensures that students need to share ideas, argue different viewpoints, provide supporting evidence for their ideas and justify their claims. This is in accordance with a sociocultural constructivist view of learning and the key roles of discussion, dialogue and

communication in the development of personal understanding (see section 1.2). The second feature is the exhibition presentation, which demands reflection and review of learning, widely recognised as important in learning and aiding the development of coherence of ideas and skills. It is worth noting, however, that these features add considerably to the difficulties of assessing the achievements of individual students.

The exhibition makes considerable demands on teachers as well as students if it is to serve its several purposes and adequately reflect all the goals and features of the programme. Although students have some role in determining the focus of the exhibition, teachers will be responsible for the eventual decision since the topic has to be shared by a group, or in some cases by the whole year group, to meet the requirement that the work is collaborative. The topic must be one that gives opportunity for students to demonstrate social skills as part of the considerable list of transdisciplinary skills and attitudes that include cooperation and empathy.

The extended timescale of the planning and completion of the exhibition means that there is opportunity for the ongoing formative assessment to be used to ensure that the intended concepts, skills and attitudes are being used and demonstrated in the activities. In regular group reviews of the progress of the exhibition both students and teachers should be using evidence to decide if what is included is enabling the learning intentions of the work to be achieved. If not, then the feedback to teachers and students can be used to align activities more closely with goals. The action that is taken through successive cycles of formative assessment should narrow gaps between the developing learning and the goals (see figure 4.1). Thus when it comes to the summative assessment of the exhibition, there should be no surprises.

The guidance given to teachers in the *Exhibition Planner* is in very general terms, such as 'There should be opportunities for students and teachers to reflect on all aspects of the exhibition throughout the process'. The reflection should be based on the standards and practices, which are also expressed in very general terms ('The exhibition reflects all major features of the programme including evidence of the five essential elements'). The aspect of formative assessment in helping students to assess their work is underlined: 'students should be aware of the criteria being used to assess performance and participation in the Exhibition.' Thus, teachers need to create the criteria and communicate them effectively to students.

Given that each exhibition will be unique, it is appropriate for the criteria to be decided as part of the planning process. In devising the criteria, teachers can benefit from the examples of teacher-devised rubrics provided on the OCC. For example, an elementary school in the USA identified criteria at four levels of performance (identified as expert, practitioner, apprentice and novice) for eight aspects of performance. Criteria are given for 'expert' and 'novice', such as the following for 'Use of transdisciplinary skills':

Expert: Students were able to apply and recognise their use of the transdisciplinary skills as indicated in their written and visual products and their journal reflections.

Novice: Students had great difficulty with five or more transdisciplinary skills as indicated in their written work, visual work, collaborative group work and research.

Another school created a matrix for criteria at four levels of meeting expectations in relation to different aspects of the work.

Such general rubrics require interpretation when applied to particular exhibition topics, which leaves the assessment of student performance open to different meanings from school to school. It raises questions as to whether the judgements that a student's work 'fully meets expectations' would be the same in school X as in School Y? Would the teacher judgement of a student's ability to 'apply and recognise the use of transdisciplinary skills' vary from one teacher to another? These questions of reliability are not addressed in the exhibition or unit assessment guidance in the PYP.

The process of judging students' achievement involves deciding how well information about their performance meets the criteria used to assess it. Whilst there is guidance for teachers in *Making the PYP Happen* about choosing tools and strategies to gather information there is little to help them apply the criteria to the information, for example, by using a 'best fit' approach (see section 5.4).

The PYP seeks to assess the exhibition with rigour and to ensure 'integrity without formally monitoring internal assessment'. When assessment tasks are unspecified as in the exhibition, experience and research suggests that rigour – and reliability of the assessed outcomes – can be best assured through having clear criteria, and some form of moderation process that gives assurance that the criteria are being consistently applied.

Two implications follow from this. The first is that, rather than schools devising their own criteria from scratch, the PYP should consider providing a set of general statements constituting a template which can be used to identify criteria appropriate to particular exhibition topics. Second is that group moderation of teachers' judgements should be recommended, where teachers discuss their judgements of students' achievements in relation to the criteria. Both of these would be included in the standards and practices for the exhibition.

# 6. Summary and key implications

## 6.1 Assessment in the PYP

Interest in the nature of formative assessment, or assessment for learning, and its potential for improving learning with any age group continues to grow around the world, notwithstanding recent calls for more robust empirical evidence to support the claims of its impact on learning. The strong emphasis given to formative assessment in the inquiry-based PYP is entirely in line with the international trend. There is solid and substantial support in the academic literature for the emphasis given to formative assessment in the PYP approach to student assessment.

However, there is evidence of a tendency for formative assessment to be used by some PYP teachers as a process that can be separated from, or loosely linked to, teaching, rather than something that is integral to effective teaching. This is an aspect that could usefully be addressed through focused professional development. There is also evidence in the PYP of some confusion between the nature and respective purposes of formative and summative assessment. This again is an issue that could be addressed through professional development.

It is useful to identify the features and principles common to all assessment, but in order to make clear the aspects of formative assessment that help learning it may be necessary to make a greater distinction in the discussion of assessment between what is formative and what is summative. For example, the criteria for effective assessment in the PYP are described as being applicable to both formative and summative assessment (IB 2007: 46). They include reference to enabling teachers to 'use scoring that is both analytical and holistic'. As noted in section 4.2, scoring is not an effective way of providing feedback to students in order to help their learning. The scores that are given to the students' work may, indeed, be used formatively in informing decisions about where they may need particular help, but the process of giving a score to summarise learning at a particular time is essentially summative.

Consideration should be given to producing a shorter list of the characteristics common to effective assessment and then lists of the particular characteristics of formative and summative assessment, drawing perhaps on Boxes 2.1 and 2.2 in section 2 of this report. The shorter list of common characteristics might be combined with a revised statement of principles of assessment (as suggested in section 3.1) and discussion of how they apply to assessment in the PYP.

Furthermore the separate description and discussion of the difference between formative and summative assessment would allow attention to be given to the different requirements for reliability and validity of assessment used for these two purposes. At present, the identification of assessment type in some of the teacher-developed overviews of

assessment (in the annotated samples of assessment in the Online Curriculum Centre) indicates some confusion about identifying assessment as formative or summative.

Although PYP practitioners do use formative assessment routinely in their classroom teaching there are indications that this predominantly takes the form of feedback from themselves to their students. It seems likely that inviting students to assess their own work and that of their peers is not common practice. While self and peer assessment might not be a realistic proposition for the youngest students, students in the upper primary school could perhaps be encouraged to engage in self- and peer-assessment at appropriate times, using shared assessment criteria. The emphasis on student reflection on their work, already present in the PYP planning frameworks, is an excellent basis for further development of this feature of formative assessment. Translated into practice, this provides opportunity for teachers to ensure that students recognise and understand the goals of their learning and participate in discussions about what they might need to do to improve aspects of their current achievement in the light of the learning goals.

The collaborative exhibition is the culmination of the PYP experience, for students, teachers and parents. Throughout their work on their exhibition topic students will have been benefiting from their teachers' guidance, their learning progress monitored through ongoing formative assessment practice and feedback. As with end-of-unit assessment, PYP teachers are required to come to summative judgements of their students' achievements when the exhibition is delivered to the audience of fellow students, teachers, parents and others. Although there is no high-stakes pressure at this stage for students or for their teachers, in relation to advancement in the IB programme, questions might nevertheless usefully be asked about the dependability of the judgements made of individual students on the basis of their exhibition performance.

Individual teachers, in consultation with their students, are free to select any study topic for the students' exhibition that is topical and motivating for the students, and which lends itself to the application of transdisciplinary skills and to collaborative research preparation and delivery. It would be difficult, given this freedom of choice, for the PYP to prescribe a common set of criteria that teachers should apply to students' evidence of learning in the exhibition when arriving at their summative judgements of individual students and of collaborative groups. This leaves teachers developing their own sets of criteria for assessing exhibition topic in question, and the scope this affords for the application of particular conceptual understanding, knowledge, skills and indeed attitudes on the part of their students.

In this context it cannot be assumed that students in different groups, classes or schools are being similarly judged against the same criteria and with the same standards of judgement. Perhaps this is not an issue at present. It might become an issue, however, if, for example, students' formally recorded PYP achievements are carried with them into the MYP and beyond, or if results are used in programme or school evaluation. The PYP should consider

developing and providing a set of general criteria statements constituting a template to be used by teachers to identify criteria appropriate to particular exhibition and unit topics.

In addition, professional development should be available to support teachers in the application of the criteria in a standardised way (as far as this is possible in practice), so that their students' achievements are as fairly assessed as possible. There is strong evidence in support of group moderation of teachers' judgements – where teachers discuss their judgements of students' achievements in relation to the assessment criteria – as an effective means not only of improving teachers' judgements but as professional development which adds to their understanding of progression in learning. At the same time, the effectiveness of moderation should be empirically evaluated periodically to ensure that the time and effort involved in implementing moderation exercises are justified, and the ultimate goal achieved.

## 6.2 Key findings and implications of the assessment review

This summary list of key implications for the PYP covers both formative and summative assessment practice:

- There is solid and substantial support in academic literature and research for the emphasis given to formative assessment in the PYP approach to student assessment.
- Formative assessment should be presented as integral to effective teaching rather than a process that is separate from, but can be linked to, teaching.
- The key features of formative and summative assessment should be clearly distinguished.
- Teachers should be encouraged to extend the range of formative assessment strategies they use beyond feedback to students.
- Attention should be given to teachers helping students to recognise the goals of their work and take part in decisions about what they need to do to achieve them.
- The PYP should consider providing a set of general criteria statements constituting a template to be used by teachers to identify criteria appropriate to particular exhibition and unit topics.
- Group moderation of teachers' judgements should be promoted, where teachers discuss their judgements of students' achievements in relation to the assessment criteria.

# References


Alexander, R. (2008) *Towards Dialogic Teaching. Rethinking Classroom Talk*. Cambridge: Dialogos.

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

Anderson, L.W. and Bourke, S.F. (2002) *Assessing Affective Characteristics in Schools,* 2nd Edition. Mahwah N.J.: Erlbaum.

ARG (Assessment Reform Group) (2002a) *Assessment for Learning: 10 Principles*. http://assessmentreformgroup.files.wordpress.com/2012/01/10principles_english.pdf

ARG (2002b) *Testing, Motivation and Learning.* Cambridge: University of Cambridge Faculty of Education. http://assessmentreformgroup.files.wordpress.com/2012/01/tml.pdf

ARG (1999) *Assessment for Learning*. Assessment Reform Group. http://www.aaia.org.uk/content/uploads/2010/06/Assessment-for-Learning-Beyond-the-Black-Box.pdf

Atkin, J.M. and Black, P.J. (2003*) Inside Science Education Reform. A history of curricular and policy change*. New York: Teachers' College Press.

Australian Academy of Science (2004 – 2012) *Primary Connections. Linking Science with Literacy.* http://science.org.au/primaryconnections/

Bennett, R.E. (2011) Formative assessment: a critical review. *Assessment in Education*, 18(1), 5-25.

Bergan, J.R., Sladeczek, I.E., Schwarz, R.D. and Smith, A.N. (1991) Effects of a measurement and planning system on kindergarteners' cognitive development and educational programming, *American Educational Research Journal,* 28(3), 683-714.

Black, P. (1993) Formative and summative assessment by teachers, *Studies in Science Education*, 21(1), 29-97.

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

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. (1998a) Assessment and classroom learning, *Assessment in Education,* 5(1), 7-74.

Black. P. and Wiliam, D. (1998b) *Inside the Black Box,* London: King's College.

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. (2012) The reliability of assessments, in J. Gardner (ed.) *Assessment and Learning*. London: Sage, 243-263.

Bliss, J. (1993) The social construction of children's scientific knowledge, in P. Black and A. Lucas (eds) *Children's Informal Ideas in Science*. London: Routledge.

Bloom, J. S., Hastings, T.J. and Madaus, G.F. (1971*) Handbook on Formative and Summative Evaluation of Students Learning.* New York: McGraw-Hill.

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.

Briggs, D.C., Ruiz-Primo, M.A., Furtak, E., Shepard, L. and Yin, Y. (2012) Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement*, 31(4), 13–17.

Brookhart, S.M. (2007) Expanding views about formative classroom assessment. A review of the literature, in J.H. McMillan (ed.), *Formative Classroom Assessment: Theory into Practice*. New York: Teachers College Press, 43-62.

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.

Bull, S., Quigley, S. and Mabbott, A. ( 2006) Computer-based formative assessment to promote reflection and learner autonomy, *Engineering Education*, 1(1), 8-18.

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), 1-14.

Care, E. and Griffin, P. (2011) Technology in assessment: Teaching and assessing skills and competencies for meeting the demands of the 21st century. ACACA Conference: Assessment for learning in the 21st century, 3-5 August 2011, Brisbane.

Carlsen, W.S. (1991) Questioning in classrooms: a sociolinguistic perspective, *Review of Educational Research*, 16(2), 157-178.

Chung, G.K.W.K. and Baker, E.L. (1997) *Technology in Action: Implications for Technology in Assessment (CSE Technical Report no 459).* Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Crooks, T.J. (1988) The impact of classroom evaluation practices on students. *Review of Educational Research*, 58(4), 438-481.

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

DfE (2011) *The Framework for the National Curriculum. A Report by the Expert Panel for the National Curriculum Review*. London: Department for Education.

Dewey, J. (1933) *How we Think: A Restatement of the Relation of Reflective Thinking to the Educative Process.* Boston, MA: D.C. Heath.

Driver, R. (1983) *The Pupil as Scientist?* Milton Keynes: Open University Press.

Driver, R., Guesne, E. and Tiberghien, A. (1985) *Children's Ideas in Science.* Milton Keynes: Open University Press.

Duschl, R.A., Schweingruber, H.A. and Shouse, A.W. (2007) *Taking Science to School: Learning and Teaching Science in Grades K-8*. Washington DC: National Research Council.

Fuchs, L.S. and Fuchs, D. (1986) Effects of systematic formative evaluation: a meta-analysis. *Exceptional Children*, 53(3), 199-208.

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

Gopnik, A., Meltzoff, A.N. and Kuhl, P. K. (1999) *The Scientist in the Crib.* New York: William Morrow.

Griffin, P., McGaw, B. and Care, E. (eds) (2012). *Assessment and Teaching of 21st Century Skills.* Dordrecht: Springer.

Griffin, P., Care, E., Bui, M. and Zoanetti, N. (2013) Development of the assessment design and delivery of collaborative problem solving in the Assessment and Teaching of 21st Century Skills project, in E. McKay (ed.), *ePedagogy in Online Learning: New Developments in Web-Mediated Human-Computer Interaction.* Hershey, PA: IGI Global.

Griffith, S.A. (2008) A proposed model for assessing quality of education. *International Review of Education*, 54(1), 99-112.

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

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

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

Harlen, W. (2013) *Assessment and Inquiry-Based Science Education: Issues of Policy and Practice.* IAP. http://www.lulu.com/content/paperback-book/assessment-inquiry-based-science-education-issues-in-policy-and-practice/13672365

Harlen, W. and Deakin Crick, R. (2003) Testing and motivation for learning, *Assessment in Education*, 10(2), 168-128.

Harlen, W. and Qualter, A. (2009) *The Teaching of Science in Primary Schools,* 4th edition. London: Routledge.

Harlen, W., Black, P. and Johnson, S. (1981) *APU Science in Schools Age 11 Report No 1*. London: HMSO.

Hattie, J. and Timperley, H. (2007) The power of feedback. *Review of Educational Research*, 77(1), 81-112.

Herl, H.E., O'Neil, H.F, Jr., Chung, G.K.W.K. and Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behaviour*, 15(3-4), 315-333.

IB (2007) *Making the PYP Happen: A Curriculum Framework for International Primary Education.* Cardiff: International Baccalaureate.

IB (2008) *Towards a Continuum of International Education*. Cardiff: International Baccalaureate.

IB (2009) *The Primary Years Programme: A basis for practice.* Cardiff: International Baccalaureate.

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

Johnson, S. (2012) *Assessing Learning in the Primary Classroom*. London: Routledge.

Johnson, S. (2013) On the reliability of high-stakes teacher assessment, *Research Papers in Education*, 18(1), 91-105.

Keogh, B., Dabell, J. and Naylor, S. (2008) *Active Assessment in English.* London: Routledge.

Kimbell, R., Wheeler, A., Miller, S. and Pollitt, A. (2009) *E-scape Portfolio Assessment Phase 3 Report*. Department for Education, Goldsmiths, University of London. http://www.gold.ac.uk/teru/projectinfo/projecttitle,5882,en.php

Kingston, N. and Nash, B. (2011) Formative assessment: A meta-analysis and a call for research. *Educational Measurement*, 30(4), 28–37.

Kulik, C. L. C. and Kulik, J. A. (1987) Mastery testing and student learning: a meta analysis, *Journal of Educational Technology Systems*, 15, 325-345.

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

Magliano, J. P. and Graesser, A. C. (2012) Computer-based assessment of student constructed responses, *Behavior Research Methods,* 44(3), 608-621.

Mansell, W., James, M and the Assessment Reform Group (2009) *Assessment in Schools. Fit for Purpose? A Commentary by the Teaching and Learning Research Programme.* London: ESRC Teaching and Learning Research Programme. http://www.tlrp.org/pub/commentaries.html

Martone, A. and Sireci, S.G. (2009) Evaluating alignment between curriculum, assessment and instruction. *Review of Educational Research*, 79(4), 1332-1361.

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

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

Mislevy, R.J. (1996) Test theory reconceived. *Journal of Educational Measurement*, 33(4): 379-416.

Natriello, G (1987) The impact of evaluation processes on students. *Educational Psychologist*, 22(2), 155-175.

Naylor, S. and Keogh, B. with Goldsworthy, A (2004) *Active Assessment in Science.* London: David Fulton.

Newton, P.E. (2007) Clarifying the purposes of educational assessment. *Assessment in Education*, 14(2), 149-170.

Newton, P.E. (2010) Educational assessment – concepts and issues: the multiple purposes of assessment, in E. Baker, B. McGaw and P. Pearson (eds.) *International Encyclopaedia of Education.* Oxford: Elsevier.

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

Next Generation Science Standards (2013) www.nextgenscience.org/next-generation-science-standards

Nichols, P.D., Meyers, J.L. and Burling, K.S. (2009) A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement*, 28(3), 14–23.

Nusche, D., Laveault, D., MacBeath, J. and Santiago, P. (2012) *OECD Reviews of Evaluation and Assessment in Education: New Zealand 201*1. Paris: OECD Publishing.

OECD (2013) *Synergies for Better Learning: an International Perspective on Evaluation and Assessment.* Paris: OECD.

OECD/CERI (2008) 21st Century Learning: Research, Innovation and Policy Directions from Recent OECD Analyses. http://www.oecd.org/site/educeri21st/40554299.pdf

O'Neil, H.F. and Klein, D.C.D. (1997) *Feasibility of Machine Scoring of Concept Maps. Technical Report 460,* University of California, Los Angeles: CRESST. www.cse.ucla.edu/products/reports/TECH460.pdf

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.

Phelan, J., Choi, K., Vendlinski, T., Baker, E. and Herman, J. (2011) Differential improvement in student understanding of mathematical principles following formative assessment intervention. *The Journal of Educational Research*, 104(5), 330-339.

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

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.

Popham, W.J. (2000) *Modern Educational Measurement: Practical Guidelines for Educational Leaders.* Needham, MA: Allyn and Bacon.

Sach, E. (2012) Teachers and testing: an investigation into teachers' perceptions of formative assessment. *Educational Studies*, 38(3), 261-276.

Sadler, D.R. (1989) Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119-144.

Schunk, D. (1996) Goals and self-evaluative influences during children's cognitive skill learning, *American Educational Research Journal,* 33(2), 359-382.

Shepard, L.A. (2009) Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement,* 28(3), 32-37.

Shermis, M. D. (2010) Automated essay scoring in a high stakes testing environment, in Shute, V.J. and Becker, B.J. (eds*) Innovative Assessment for the 21st Century.* Dordrecht: Springer.

Shute, V.J. (2008) Focus on formative feedback. *Review of Educational Research,* 78(1), 153-189.

Shute, V.J. (2011) Stealth assessment in computer-based games to support learning, *Computer Games and Instruction,* 55(2), 503-521.

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

SQA (2009) *Guide to Assessment*. Glasgow: Scottish Qualifications Authority. http://www.sqa.org.uk/sqa/controller?p_service=Front.search&pContentID=41454&q=guide%20to%20assessment

STA (2012) *2013 Early Years Foundation Stage Profile Handbook*. London: Standards and Testing Agency.

Stiggins R.J. (2001) *Student-Involved Classroom Assessment.* Upper Saddle River, New Jersey: Merrill Prentice Hall.

Stobart, G. (2008) *Testing Times. The Uses and Abuses of Assessment.* London: Routledge.

Tan, L. and Bibby,Y. (2011) Performance Comparison between IB School Students and Non-IB School Students on the International Schools' Assessment (ISA) and on the Social and Emotional Wellbeing Questionnaire. Report for IB from ACER http://www.ibo.org/research/policy/programmevalidation/pyp/documents/IB_ISA_report_Nov2011_Final.pdf

Tincani, M. and Crozier, S. (2007) Comparing brief and extended wail-time in small group instruction for children with challenging behaviour. *Journal of Behavioral Education*, 16(4), 355-367.

Torrance, H. (2007) Assessment *as* learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education,* 14(3), 281-294.

Torrance, H. (2011) Using assessment to drive the reform of schooling: Time to stop pursuing the chimera? *Educational Studies*, 59(4), 459-485.

Tunstall, P. and Gipps, C. (1996a) "How does your teacher help you to make your work better?" Children's understanding of formative assessment. *The Curriculum Journal,* 7(2), 185-203.

Tunstall, P and Gipps, C. (1996b) Teacher feedback to young children in formative assessment: a typology. *British Educational Research Journal*, 22(4), 389-404.

Tyler, R., Gagné, R.M. and Scriven, M. (1967) *Perspectives of Curriculum Evaluation.* American Educational Research Association Monograph Series on Curriculum Evaluation. Chicago: Rand McNally.

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

White, B.Y. and Frederiksen, J.T. (1998) Inquiry, modelling and metacognition: making science accessible to all students. *Cognition and Instruction,* 16(1), 3-118.

Wiliam, D. (1993) Reconceptualising validity, dependability and reliability for National Curriculum Assessment. Paper given at the British Educational Research Association conference, September 1993.

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

Wiliam, D (2010) The role of formative assessment in effective learning environments, in H.Dumont, D. Istance and F. Benavides (eds) *The Nature of Learning: Using Research to Inspire Practice*. Paris: OECD, 135-159.

Wiliam, D. (2011) *Embedded Formative Assessment*. Bloomington, IN: Solution Tree Press.

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.

WNCP (Western and Northern Canadian Protocol) (2006) *Rethinking Classroom Assessment with Purpose in Mind*. Manitoba Education, Citizenship and Youth. http://www.edu.gov.mb.ca/k12/assess/wncp/rethinking_assess_mb.pdf