Chapter 3
Exploring Relations Between Formative and Summative Assessment

Jens Dolin, Paul Black, Wynne Harlen, and Andrée Tiberghien

The goals of STEM education go beyond specific knowledge and skills. They include a wider range of competencies and contribute, as does work in other domains, to what are variously described as twenty-first-century skills or life skills necessary for personal fulfilment, employment, citizenship and social responsibility. For example, the countries of the European Union have agreed upon a framework of key competencies that are being reflected in the school curricula of most of these countries (European Commission 2011). They include competencies in communication, mathematics, science and technology, learning to learn, social and civic responsibility, initiative and cultural awareness. Similarly, in response to the rapid changes in technology that make information readily available, the countries of the OECD are emphasising the importance of competencies relating to selecting, synthesising, transforming and applying information, thinking creatively, collaborating with others and communicating effectively (OECD 2013).

These general educational trends have influenced the STEM learning goals (as described in Chap. 1). The generic competencies have to be learned while working with subject-specific content and problems, changing both curriculum and pedagogy. Simultaneously, the STEM domains have developed their curriculum goals and descriptions to focus on some domain-specific competencies, like modelling and investigating.

J. Dolin (✉)
Department of Science Education, University of Copenhagen, Copenhagen, Denmark
e-mail: dolin@ind.ku.dk

P. Black
School of Education, Communication and Society, King's College, London, UK

W. Harlen
Graduate School of Education, University of Bristol, Bristol, UK

A. Tiberghien
University of Lyon and Centre National de la Recherche Scientifique, ICAR (CNRS), Lyon, France

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science Education Research 4, DOI 10.1007/978-3-319-63248-3_3

Achieving these new goals may require changes to be made, not only in curriculum content and pedagogy but also in assessment content and tools, given the well-established impact of what is assessed on what is taught and how it is taught (Harlen and Deakin Crick 2003). The central goal is to improve learning, and assessment is a tool to help achieve this goal; however, unless the full range of goals is assessed, those goals not included are likely to be given less attention. As a result, developing new competencies may well remain as an aspiration on paper unless appropriate assessment tools and processes are developed. Developing assessment approaches able to capture these new goals is therefore necessary and at the heart of many recent projects such as SAILS (Strategies for Assessment of Inquiry Learning in Science – an EU FP7 project), TAPS (Teacher Assessment in Primary Science – a Primary Science Teaching Trust project) and the ASSIST-ME project. The ASSIST-ME project has researched the implementation in the partner countries of the formative use of four assessment methods able to capture inquiry processes within STM subjects. Ideally, the high-stakes assessment that influences teaching practice should use similar approaches. As it was beyond the scope of ASSIST-ME to determine national examinations, the project pragmatically investigated how the formative use of the four assessment methods are in alignment with the more high-stakes assessments, especially the final examinations, and how current high-stakes assessments should be changed to promote effective formative assessment.

This chapter sets these investigations into a theoretical frame by characterising two key purposes of assessment, formative and summative, and considering how they are related and can be brought together in developing a dependable approach to summative assessment using evidence collected and used in formative assessment. The third purpose of assessment, accountability, is dealt with as a special use of summative assessment. The range of approaches to using assessment formatively – to influence learning activities as they take place – and summatively, to record evidence of what has been learned at certain times, is discussed. Some examples from the ASSIST-ME project illustrate the variety of approaches to assessment and the overlapping relations between formative and summative use of assessment.

The Characteristics of Formative and Summative Assessment

The distinction between formative and summative as applied to student assessment emerged in the 1960s, having originated in identifying the roles of programme evaluation in the development of new curriculum materials. Evaluation described as formative was conducted during the process of materials development to provide information about how to revise early drafts, while summative evaluation usually provided some measure of the effectiveness of the final draft. So, in the case of student learning, the main purpose of formative assessment is seen as helping learning,

while the main purpose of summative assessment is to provide information about what learning has been achieved at a certain time. Even major texts on pedagogy have until recently paid almost no attention to the potential of formative assessment to assist the day-to-day learning of students (Black 2016).

Before we can discuss how to combine aspects of assessment for formative and summative purposes, it is helpful to define and identify the characteristics they have in common and their differences. It is important to recognise that formative and summative refer to different *purposes* of assessment and not to different kinds or *forms* of assessment. As we will see later, data of the same kind can, in some circumstances, be used both formatively and summatively. It is the impact of assessment in a particular instance and the use made of the data that identifies the assessment as formative or summative and not the form of the data. This distinction is more easily and more often overlooked in the case of formative assessment, so we begin by looking at the characteristics of assessment used to enhance learning. However, in recognition of the way assessment language is commonly used in practice, we will often use the phrase 'formative assessment' for the more correct 'formative purpose of assessment' or 'formative use of assessment evidence'.

Formative Assessment

Although the concept of using assessment to identify difficulties in learning and to support growth in learning has been embraced in many countries across the world, there is considerable diversity in how assessment is translated into practice and conflicting views about how it is defined. In particular there is a division, to which Bennett (2011) draws attention, between 'those who believe formative assessment refers to an instrument (e.g., Pearson 2005), as in a diagnostic test, an 'interim' assessment, or an item bank from which teachers might create those tests' (Bennett 2011, p. 6) and those who view it as a process. There are considerable practical and implementation differences between these two views. Expressed rather starkly, the 'instrument' view involves the use of a tool such as a test or task that will yield information about current competencies, often in the form of a score or judgement. Seen in this way, formative assessment is much less dependent on the ability of teachers to ask appropriate questions or make relevant observations than is the case for the process view, which produces 'a qualitative insight into students learning rather than a score' (Shepherd 2008). Of course, this exaggerates the differences between these views, and in practice there is overlap between the two positions of how formative assessment is conceived and practised. Nevertheless, the most widely used definitions of formative assessment emphasise the process view, for instance:

> *Formative assessment is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there.* (ARG 2002)

> *Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes.* (CCSSE, in McManus 2008)

It can in these definitions be assumed that the assessment activities make use of a variety of methods (which could be seen as instruments) to gain insight into present levels of student competence, and the process is very much about using this method to get access to students' learning. The ARG definition focuses on the learner and the learner's learning process, while the CCSSE definition embeds formative assessment in instruction in order to improve both students' learning and teachers' teaching.

The focus on process is underpinned by an individual or sociocultural constructivist (see later) concept of learning, that is, one that views learning as making sense of new experience starting from existing ideas and competencies. Evidence of what students already know and can do, in relation to short-term goals of a particular lesson or group of lessons on a topic, informs decisions about the steps needed to make progress. The process involves teachers and students in:

- Establishing clear goals, and progression steps (including criteria) towards them, that are shared between teacher and students
- Designing or choosing a method for collecting data about student performance
- Using the method for gathering evidence of where students are in relation to these goals
- Interpreting this evidence to decide what is needed to help students towards the goals
- Deciding what action can help the student through the next progression step

In Fig. 3.1 these actions are connected by arrows into a repeated cycle of events (the arrows leading away from the cycle to the right are explained later). A, B and C indicate activities that lead students to work towards the goals. Evidence gathered in activity A is interpreted in relation to where students are in relation to the goals. This information facilitates decisions about appropriate next steps and how to take them, leading to activity B. The cycle is repeated, and the effects of decisions at one time are assessed at a later time as part of the ongoing process. How often this happens depends on the nature of the goals; a cycle might take place over several lessons or might take only a few minutes so that it is carried out several times in one lesson.

In practice the process may not be as formal and teacher directed as it appears in this description. The actions are not 'stages' in a lesson; rather, the cycle represents a framework for helping teachers and students to become more conscious of what is involved in learning and using this to help further learning and also to adapt teaching. The learners are at the centre of the process, and the two-headed arrows indicate their role as both providers of evidence and receivers of information to improve their learning. This gives students a role in their own assessment that helps them to come to understand the process of learning, to work towards explicit goals and standards and to modify what they do in relation to constructive task-related feedback

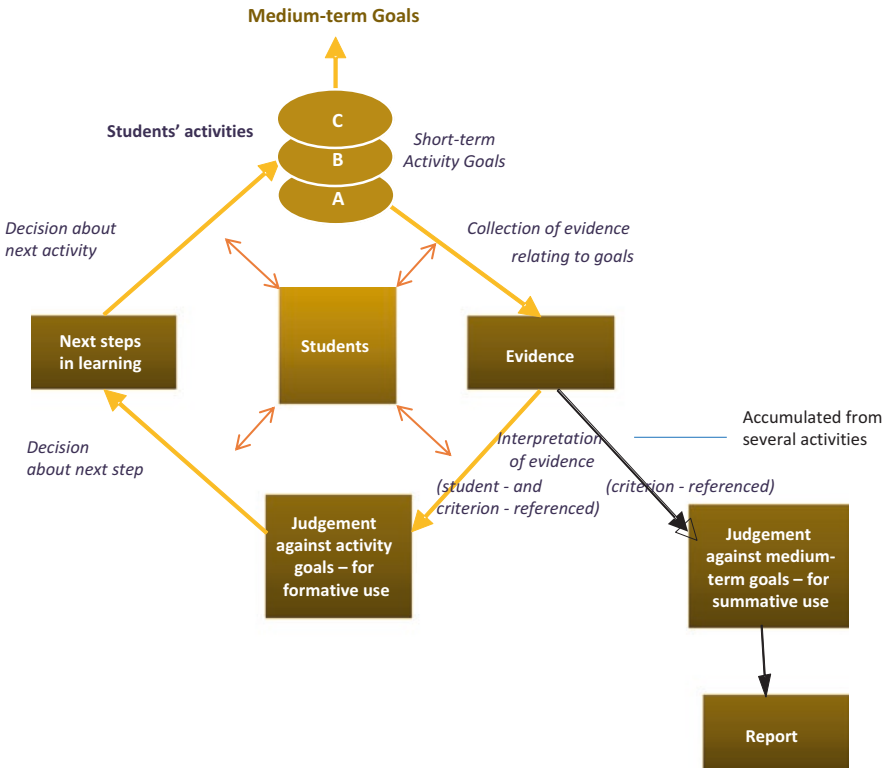

Fig. 3.1 The assessment process for both summative and formative purposes

from teachers and from other students and even from themselves. Besides enhancing learning, this involvement of the student in his or her learning process also engages and motivates students. This will further enhance learning and also guide students to a more self-directed learning (Wiliam 2011). Thus, activating students as instructional resources for one another will often be an important part of the teacher's task.

Parts of the formative cycle can be carried out in different ways depending on the type of activity and the age and ability of the students.

Ways in which evidence can be collected include:

- Observing students as they work, asking questions to probe their understanding, listening to their explanations and engaging in dialogue (ephemeral evidence often collected 'on-the-fly')
- Studying the outcome of their work, such as drawings, video and reports
- Embedding special tasks designed to require particular ideas and competencies in regular work
- Giving tests or quizzes (teacher made or externally produced)

Chapters 4, 5, 6 and 7 in this book describe different ways of collecting and judging evidence of student learning. The evidence gathered is judged in terms of what it indicates about existing ideas and competencies required to meet the lesson goals. The judgement may be made:

- By reference to a fixed set of predetermined criteria which provide concise descriptions of what students are expected to know and be able to do at a specific stage of their education (criterion-referenced)
- By reference to what is usual for similar students to be able to do – presupposing a normal distribution, that is, a bell-shaped curve, of students' performance (norm-referenced)
- By reference to expectations of the performance of particular students, taking into account what the student was previously able to do and the progress that the student has made over time (student-referenced or ipsative)
- A mixture of these

Judgement of the evidence in terms of progress towards the learning goals provides information to enable the teacher and students to decide what next steps are needed or, indeed, whether no immediate new action is needed if an activity is proceeding productively. Note that a distinction is made here between *evidence* (what was said, written or created) and *judgement* (interpreting what this means for the degree of goal attainment or for progress of individual or groups of students). This distinction is relevant to using assessment data for formative and summative purposes.

In formative assessment, taking the next steps involves feedback into the teaching-learning process. Feedback here, as in other situations, means giving responses to a product or process or event to improve performance. In the context of teaching, consistent with the social constructivist view of learning, feedback is from student to teacher, from teacher to student or from student to student. Feedback from student to teacher enables the teacher to know how students are responding to the learning activities and enables the teacher to know what action to take to adjust the opportunities and challenges provided to students. Such feedback may be used by teachers to adjust teaching in future planning of similar activities for other students. Feedback from teachers to students should give students information about how they can improve their work or take their learning forward. Just giving marks or grades that only indicate how well the work is judged to be is not consistent with the aim of using assessment to help learning. A controlled experimental study by Butler (1987, 1988) found that giving marks, with or without added comments, was less effective in improving students' work than giving comments only, while the extensive research studies of Dweck (2000) show that feedback that includes marks can have long-term negative effects on students' beliefs about their own ability to learn. What seems to happen is that students are looking for judgements rather than help with further learning; they seize upon marks and ignore any accompanying comments. Student to student feedback (peer feedback) may take place in the course

of dialogue among students when they discuss the strengths and weaknesses of one another's work (see Chap. 6).

This 'constructive' use of formative assessment hinges on the ability of the teacher (or another provider of feedback) to actually give recommendations that are relevant and effective for improvement. Many of the findings in chapter four to seven point at this as the key issue in the implementation of formative assessment methods. The teacher needs to have a well-established pedagogical content knowledge, and the students need to learn how to give and receive feedback.

A Rationale for Formative Assessment

Formative assessment, unlike summative assessment, is not always a formal requirement of a teacher's role. Teachers generally have the duty to conduct some form of assessment in order to keep records of students' achievement and to report, for instance, regularly to parents. In some of the countries participating in ASSIST-ME, the formal requirements of teaching seemed more about judgement than about learning. If there is no similar imposed requirement for formative assessment, teachers and others may need to be persuaded of its value if they are to make the effort to include it in their practice.

Across the countries participating in ASSIST-ME, the hindrance most often expressed by teachers using formative assessment in their teaching was lack of time. Having been introduced to various assessment methods and been working with the implementation for formative purpose for some time, all teachers agreed upon the usefulness and effectiveness of formative assessment. They also agreed that students were satisfied and that learning was enhanced. But 'how do we get the time for it?' was a common question.

The teachers often perceived formative assessment as an add-on to the teaching, something extra instead of understanding it as a central and integrated part of teaching and learning. And in a way, they were right: learning takes time. Teachers' perception of not having time for formative assessment might simply be an illustration of the long-known fact that most curricula are overloaded, and in order to cover it, the focus is more on teaching than learning. But it might also be an illustration of what you choose to spend time on relates to what you think matters.

An important rationale for formative assessment practices follows from current perceptions of how learning takes place. For some time, it has been recognised that learners have an active role in constructing their understanding; it is not something that can be received ready-made from others, as in the more simple versions of the theory of learning described as *behaviourist.* Rather, we recognise that developing understanding requires active participation of learners in constructing their learning. This accords with a *cognitive constructivist* view of learning, that is, learners making sense of new experiences starting from existing ideas and competencies.

Recognition that learning is not entirely, or even mainly, an individual matter but takes place through social interaction is the basis of the *sociocultural constructivist* perspective of learning. In this view, understanding results from making sense of new experience with others rather than by working individually. In group situations the individual learner takes from a shared experience what is needed to help his or her understanding (internalises) and then communicates the result as an input into the group discussion (externalises). There is a constant to-ing and fro-ing from individual to group as knowledge is constructed communally through social interaction and dialogue. Physical resources and language also have important roles to play (James 2012). 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).

From this perspective, learning is conceptualised as a social and collaborative activity in which people develop their thinking together. In classrooms the interaction between students is mostly face-to-face, but learning from and with others can also be through the written word. Feedback to students in writing can be an effective channel for dialogue between teacher and students, providing that the comments take students' learning forward and the students have time to read and reflect on the comments and perhaps make amendments or additions to their work in response to these comments. Thus we can promote progress towards learning goals by teachers in a range of ways, such as encouraging collaboration and group work, providing clear goals, giving feedback, enhancing dialogue and negotiating. Approaches such as these contribute to the process of formative assessment.

The Role of Learning Progression

A clearer definition of formative assessment and description of what it involves in practice is needed to establish sound formative practice. It was a common finding among the researchers involved in ASSIST-ME that teachers in general have a vague and implicit understanding and use of formative assessment. Many teachers didn't distinguish between formative and summative use of assessment, and they often used quite weak feedback practices. In particular, working with learning progression steps as a central prerequisite for feedback processes was quite new for many teachers.

A learning progression describes 'successively more sophisticated ways of reasoning within a content domain that follow one another as students learn' (Smith et al. 2006, p. 1, cited in Duncan and Hmelo-Silver 2009, p. 606). Basically, a learning progression is necessary to guide planning for feedback that seeks to identify and build towards next steps in learning. It is also crucial if students are to be involved in the formative process. Broad learning goals must be broken down into sub-goals with success criteria related to each sub-goal. Traditionally, teachers have not made these steps explicit. The steps were inside their heads, as they said when interviewed, as part of their professional knowledge. This works fine for teachers

giving relatively informal feedback to students, but if students are to be involved, by self-assessment or peer assessment, the steps need to be explicit, and students need to know what they are. More structured feedback is also necessary if the formative assessment is to have a certain degree of reliability. Very often an increase in mastering a given area of knowledge or of a competence is described using words from a generic taxonomy like Blooms taxonomy or the SOLO taxonomy. This is, for example, the case in descriptions of most grading scales and curriculum statements. But when teachers in the project reflected on their students' learning path within a specific part of the discipline, based on their experience as teachers, they did not construct competence levels following a standard taxonomy across disciplines. Very often they built a sequence of building blocks, more like stepping stones spread over a muddy field where you have to stand on each stone in order to have covered the whole field. Accordingly, they found that they would not be able to reuse learning progressions from one discipline area to another but needed to design them independently for each new content. This is in accordance with the findings from a major science education project working with learning progressions (Alonzo and Gotwals 2012).

Teachers in the Local Working Groups (LWGs) found it time-consuming and hard work to make levels of attainment explicit to the students. They acknowledged the fact that it is valuable for teaching and for the students, to make explicit what they normally do without conscious thought, but it was simply too time-consuming. Also, some teachers experience a need to work with colleagues on formulating learning progressions, and there was not always time to do this.

As expressed by two Danish science teachers involved in ASSIST-ME:

> There is a big difference between working with learning progression and to be aware of learning progression. I do not work much with it but I am very conscious of it. I do think progression clearly into the way I structure my lessons and how I ask students, but I have never before this project made it clear for the students that we have this learning progression.

> I would normally never write progression steps to myself. We know them; they are deep inside of us. So, it should only be for the benefit of the students, to make the demands clear for them.

Most teachers in the project did not have clear criteria or did not make them explicit for the students, and they did not see (and did not use) formative assessment as a specific procedure with specific quality characteristics. This has serious implications for both the validity and especially the reliability of their judgements and hence the potential summative use of the formative processes.

Summative Assessment

The aim of summative assessment is generally to report on students' level of learning at a particular time, rather than to impact on ongoing learning, as in the case of formative assessment. In some cases an assessment can, to an extent, serve both

purposes, as we discuss later, but first it is helpful to distinguish the separate characteristics of summative assessment.

Assessment for summative purposes involves collecting, interpreting and reporting evidence of learning. Interpretation of evidence is in relation to the goals that students are intended to have achieved at a certain point, such as the end of a year, semester or stage. These are goals that can be described as medium-term, as distinct from the short-term goals of particular lessons or topics and from long-term goals such as 'big' ideas which are achieved over the whole period of school education.

There are several different ways of collecting evidence for summative assessment: by administering tests or examinations, summarising observations and records kept during the time over which learning is being reported, creating a portfolio of work, embedding special tasks in regular activities, engaging in computer-based tasks or some combination of these. When choosing how to collect evidence, consideration has to be given to the uses to be made of the data. Uses vary from routine reporting of the achievement of individual students to parents, to other teachers and to students themselves, to keeping records of the performance of groups of students by age, gender, background, etc. Some of these uses have 'high stakes' for teachers, schools and students themselves, in that they are used for making evaluative judgements which affect student selection, students' own decisions, school reputation and placement on league tables and even funding. Particular uses have implications for the degree of validity and reliability of the data used. It is important, therefore, to understand the concepts of validity and reliability and their interactions with each other.

Validity and Reliability

All assessment involves the generation, interpretation and communication of data (Harlen 2013). The same processes are involved whether the purpose is primarily formative or summative, and the main purpose of these processes is to provide inferences about the knowledge, skills and competencies that students possess. It is the way in which these processes are carried out, and the way that inferences are drawn that determine the quality of an assessment. Assessment quality is generally described in terms of two concepts – validity and reliability. Superficially we could say that validity has to do with the 'what' and 'how' of assessment, while reliability has to do with 'how well' (Johnson 2012). But there is far more to these concepts than this, as we now see.

Validity

It is usual to define validity of an assessment in terms of how well what is assessed corresponds with the behaviour or learning outcomes that it is intended should be assessed, i.e. about which evidence is required. Determining the extent to which this

is the case is complex and involves judgements of various kinds. Different types of validity have been proposed depending on the kind of information used in judging the validity. For instance, *content validity* refers to how adequately the assessment covers the subject domain being taught and is usually based on the judgement of experts in the subject. However, content coverage is not enough to describe the full range of a test or other assessment tool. The important requirement is that the assessment samples all those aspects – but only those aspects – of students' achievement relevant to the particular purpose of the assessment. Including irrelevant aspects (construct irrelevance) is as much a threat to validity as omitting relevant aspects (construct under-representation) (Stobart 2008). An example of construct irrelevance could be that inquiry-based science education problems are often broadly described in real-life settings – demanding good reading skills. So, these problems assess reading competence as well as science competence. This gives a biased result – unless, that is, the construct assessed includes reading of, for instance, social scientific issues as part of the learning demands. This is why *construct validity* is an increasingly prevalent validity concept, subsuming many of the other validity measures. Construct validity is based on a description of the skills and competences to be assessed and their relations (a so-called framework or construct). It could be a theoretical understanding or a model of a competence. A test will then have construct validity if it is able to assess the underlying construct of competence.

Another form of validity, *consequential validity*, is not a property of the assessment instrument or procedure itself but is a judgement of how appropriate the assessment results are for the uses to which they are put. The validity of an assessment tool is reduced if inferences drawn on the basis of the results are not justified. 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. On a more general level, a test may say something about student's knowledge within a specific area of knowledge but is not a valid basis for predicting the student's study success in further education.

So, validity is not a property of an assessment method or instrument regardless of the circumstances in which it is used. This situation is formally expressed in the definition by Messick (1989, p. 13) of validity as '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'.

Reliability

The reliability of an assessment refers to the extent to which the results can be said to be of acceptable consistency or accuracy for a particular use. There are many factors that can reduce the reliability of an assessment. For example, reliability is reduced if the outcomes are dependent on who conducts the assessment (which

teacher or oral examiner), on who rates the students' assessment performances (which scorer or observer or external examiner at oral examinations) or on the particular questions used in a written test when these can only test a sample of all the different topics and levels of learning included in the curriculum being tested. 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. Thus reliability is often defined as, and measured by, the extent to which the assessment, if repeated, would give the same result. For summative assessments which are conducted on a national scale and which have important consequences for all involved, extensive and sophisticated psychometric analyses have been used to explore their reliability (Baird and Black 2013). The most common used is Cronbach's α, measuring internal consistency, and Cohen's κ, measuring degree of agreement between two observers.

Reliability is important in relation to both formative and summative assessment. In the case of formative assessment, reliability is often lower, and it matters less than for summative assessment. This is because the notion of making a repeatable judgement and treating all students in the same way is not equally relevant when the purpose is to support well-founded decisions about next steps for individual students who may be at different stages in their learning and require different kinds of feedback. But even if the judgement of specific evidence should be student sensitive, the interpretation of student evidence must be based on the same criteria, using the same indicators. This is important because the judgements that constitute the field at hand form the landscape of meanings, definitions and importance in which the students should find themselves (Dahler-Larsen 2014). If formative assessment is systematically flawed, that is a problem because it could give students a wrong image of the field they have to learn and the demands they have to live up to. The reason why we sometimes live with less reliable formative assessment is that it is local and not systematically biased in a large scale. But to have the potential to be used for summative purposes, also the formative judgements need to be reliable.

Validity and Reliability Interactions

It follows from the definition of reliability that tests comprising questions where students choose between fixed alternative answers (multiple choice), that can be machine marked, are more reliable than ones that require answers to be created and that require some judgement in the scoring, such as open response format tests. However, the latter may be a more valid assessment if the purpose is to find out what answers students can produce rather than asking them to choose from specified alternative answers. It has also been shown that if students are asked, in a multiple-choice test, both to choose the best response and to explain their choice, many who choose the correct response are unable to give a valid reason for that response (Osborne et al. 2016).

Although it is desirable for summative assessment to be both highly reliable and valid, in practice there is a limit to optimising both. Normally, there is a conflict between high reliability and high validity, so that it is necessary in practice to adopt a compromise between them. An assessment which is highly reliable (like a simple multiple-choice test) is often low in validity. This is especially true for assessment of competencies and more advanced skills. Assessing complex demands validly is complicated and time-consuming, meaning that it is difficult to do in a reliable way. However, if an assessment is low in validity, it can have damaging effects, because those using the results will make incorrect inferences about the learner's capability. So, for example, students may infer, from a strong result, that they will succeed in advanced study of a subject but may discover when committed to that subject that its demands are radically different from those which they had been able to meet in the invalid summative assessments.

One aim of ASSIST-ME was to investigate these relations between formative and summative use of assessment in order to produce guidelines for changing the summative use of assessment so that it would be in alignment with the formative use. A key issue is that if the same evidence used formatively is also going to be used summatively, it needs to be (re)interpreted reliably. This means that the assessment should fulfil some psychometric measures for reliability often linked to summative assessment, like the Cronbach α for internal consistency. Here, the distinction between *evidence* and *judgement*, mentioned earlier, becomes important. For formative assessment, the evidence is interpreted both in relation to the goals (criterion-referenced) and in relation to the progress of a student towards the goals of a particular lesson or sequence of lessons, the next steps relating to where a student has reached and knowledge of the student's capabilities (student-referenced). For summative assessment, the evidence collected over a period of time needs to be judged only in relation to more general, or medium-term, goals (criterion-referenced) that apply over that period of time and to all students. We will later take up further what this means in practice, but it is relevant to note here that this dual use of evidence involves teachers in setting up longer-term goals and working towards them through breaking down the longer-term goals into sub-goals for lessons or sequences of lessons with clear goals and criteria for each sequence. This turned out to be one of the key challenges for the teachers participating in the ASSIST-ME project. As previously identified, working with implementing formative uses of assessment very quickly turned out to be a question of making learning progressions explicit within the actual subject matter and steering the teaching/learning processes in the light of these progressions. The ASSIST-ME teachers were struggling, together with the researchers, in formulating learning goals. The teachers were provided with a template with space for sub-goals and a three-step progression for each sub-goal and a list of verbs from the SOLO taxonomy and Bloom's taxonomy. But reality turned for most teachers out to be more complicated, as two Danish science teachers explained:

> It is difficult to see which progression steps fit the students' actual performance – at the end of the day it will always be an interpretation with a lot of uncertainty. You can 'explain' or 'discuss' on many levels.

> The same generic taxonomy will be different from subject to subject – the one formulated in biology may not be usable in mathematics. 'Explain' can be on a very high level in mathematics while it might be low in other subjects.

The generic taxonomies did not fit the classroom. The categories were too abstract and too far from subject practice. As mentioned earlier, the teachers found that their teaching (and students' learning) did not follow a path described via a generic taxonomy, for instance, like climbing a ladder step by step beginning at the first step. Rather, it followed a subject-specific logic that has evolved through the teachers' professional life. It could be described as crossing a river using stepping stones spread across the river bed or putting a jigsaw together. This made it quite time-consuming to establish the progression steps needed for the formative feedback processes – they had to be constructed for every subject sequence.

More problematic than the time it took, was the influence it had on the teaching, the whole structure of the lessons.

> (Progression) templates to use in the daily teaching have to be very precise and you need one for almost every lesson or sequence, so, you risk to lose the overview. You focus on individual issues and lose the whole picture. … If you give this very comprehensive material to students they will give up. And it has to be very concrete – students can't use very general concepts.

> The more systematically you do it, the less freedom you have.

> I find it unfortunate if you teach with the progression steps totally described in any detail. You lose something – it tends to be about performance instead of learning. The more detailed, the more focus on: Now the two and the five competences are achieved. If we constantly focus on how they can assess themselves, then we remove the curiosity and the joy and focus on what you have to perform.

> From a basic human point of view something inside me is against this visible learning, where you represent people as rational thinking beings. I don't think this is the whole story. I more think that it is a question of curiosity and something quite impalpable. We lose a lot by doing this – and what about the Bildung – where is that in the progression plan?

Some teachers, though, were more positive and pointed at the fact that the progression steps describe something we want students to learn. They are often important competencies and through visible learning help students steer their own learning processes. But it could be a high price to pay to obtain this. Students risk to be pushed towards a more performance-oriented regime at the cost of their mastery of the subject (Midgley et al. 2001). It turned out to be a very delicate balance.

Issues in Using Tests

Tests are often the method of choice for gathering information for summative assessment on the grounds of 'fairness,' since they appear to treat all students in the same way. However, giving all students the same task is not the same as giving them equal opportunities to show what they understand or can do. This becomes clear

when looking at test items. The subject matter used to pose a question, the language used, the amount of reading and writing involved and familiarity with the test situation are among the many factors that will advantage some students and disadvantage others. Research clearly shows how students' performance is influenced by the situation or context in which the task is set – students perform differently in different situations and especially low performers benefit from rich, authentic assessment situations (Dolin and Krogh 2010). Other well-known problems with tests stem from the necessarily limited number of items that can be included, meaning that, as pointed out above, a selection of contexts and problems has to be made and that a different selection is likely to lead to different results.

Further, when test results are used for high-stakes judgements, of students, teachers or schools, the tendency to 'teach to the test' in striving for high scores has a narrowing impact on the curriculum content and on pedagogy. Moreover, the validity of the assessment is threatened when there is a stark contrast between the mode of learning in inquiry-based activities and the mode of testing. Students, who are used to working in groups, sharing ideas and reviewing their own work in relation to comments and reactions of others, may find it difficult to express their understanding and their acquired competencies in a conventional test or examination setting.

Many of the most severe problems of tests arise from the use made of the results. This is an important factor in determining the degree of stress that students may feel and indeed may influence the perception of the amount of testing they experience. For example, when test results are used for high stakes, where decisions affect students' opportunities (as in the end of primary tests in some countries with highly selective secondary education and end of high school examination in others), students, and their parents, know that the results matter and become stressed by the process. Teachers might add to this stress by giving repeated practice tests. However, stress can also become an issue when the tests matter a great deal for the school although not for the individual students, as in the case of the national tests in England at the end of primary school (which are entirely about accountability). Insofar as teachers are stressed about the results, they may transfer this stress to students, spending considerable time directly preparing students by creating more tests for practice. This adds to the experience of students of school work as being dominated by testing. It can be argued that it is the use rather than the nature of testing that is at fault. However, until the policy that leads to this use can be changed, it is unfortunately difficult for the potential of tests to provide useful guidance to teacher and students to be realised.

The overarching criterion, by which any combination of assessment methods is to be judged, is the degree to which they both support and reflect student learning. Such learning has different dimensions, ranging from learning through interactive dialogue to learning to tackle realistic complex tasks which require the integration of different elements of one's learning (Black 2016, p.733f).

Formative and Summative Assessment: Dimension or Dichotomy?

As the examples from the project in Chaps. 4, 5, 6 and 7 show, there is considerable diversity in the ways in which formative and summative assessment is practised. For instance, feedback can be offered to students as very informal and unstructured comments, 'on-the-fly', or more formal written comments on a structured piece of work or classroom test. Some of the features of these events, such as the use of tests, may seem – at least superficially, until their use is examined more carefully – closer to a summative than a formative purpose. This suggests that the relationship between formative and summative assessment might be better described as a dimension rather than a dichotomy as suggested in Fig. 3.2 (based on Harlen 2012).

At the extremes are the practices and uses that most clearly distinguish between assessment for learning and assessment of learning. At the purely formative end,

Formative<----------->Summative				
	Informal formative	Formal formative	Informal summative	Formal summative
Major focus	What are the next steps in learning?		What has been achieved to date?	
Purpose	To inform next steps in teaching and learning	To inform next steps in teaching and learning	To monitor progress against plans	To record achievements of individuals
How evidence collected	As normal part of class work	Introduced into normal class work	Introduced into normal class work	Separate task or test
Basis of judgement	Student- and criterion-referenced	Student and criterion-referenced	Criterion and student-referenced	Criterion-referenced
Judged by	Student and teacher	Teacher and student	Teacher	Teacher or external examiner
Action taken	Feedback to students and teacher	Feedback to students and into teaching plans	Feedback to students and into teaching plans	Report to student, parent, other teachers, etc.
Epithet	Assessment for learning	Matching	Dip stick	Assessment of learning
Examples of feedback modes	Verbal feedback on-the-fly	Written feedback on classroom work	Response to informal test or quiz	Synoptic report on achievement of course goals

Fig. 3.2 Dimensions of assessment purposes and practices

assessment is integral to student-teacher interaction and is also part of the student's role. The teacher and student consider work in relation to the activity goals that are appropriate for the particular learner, and so the judgements are essentially student-referenced. The central purpose is to enable teacher and students to identify next steps in learning and to know how to take these. At the purely summative end of the dimension, the purpose is to give an account of what has been achieved at certain points. For this purpose, assessment should result in a dependable report on the achievements of each individual student. Although self-assessment may be part of the process, the ultimate responsibility for giving a fair account of how each student's learning compares with the intended learning goals should rest with the teacher, whether or not the evidence is obtained by external tests or by judgements of regular work.

Between these extremes, it is possible to identify a range of procedures having various roles in teaching and learning. Indeed, there is in practice considerable overlap across the vertical boundaries inside the above table – they are not rigid barriers. For instance, many teachers would begin a new topic by finding out what the students already know, the purpose being to inform the teaching plans and maybe to identify the point of development of each individual. Similarly, at the end of a section of work, teachers often give an informal test to assess whether new ideas have been grasped or need consolidation.

Within both the more formative and more summative parts of the dimension, there are considerable variations. Classroom observations of teachers' formative assessment practices by Cowie and Bell (1999) were interpreted as indicating two forms of formative assessment: 'planned' and 'interactive'. Planned formative assessment concerns the whole class, and the teacher's purpose is to find out how far the learning has progressed in relation to the lesson goals. Information gathering, perhaps by giving a brief class test or special task, is planned and prepared ahead; the findings are fed back into teaching. This is similar, in many ways, to 'informal summative'. The Structured Assessment Dialogue described in Chap. 5 is an example of such a tightly structured assessment method that can be used for both formative and summative purposes without further adaptation. 'Interactive' formative assessment is not planned ahead in this way; it arises from the learning activity (on-the-fly). Its function is to help the learning of individuals or groups; feedback is both to the teacher and the learners and is immediate. It has the attributes of 'informal formative'. In practice, the two forms are difficult to distinguish. A teacher might ask an open question and encourage interactive dialogue across the whole class – the interaction is planned, in part, because the teacher might believe that engaging in dialogue is itself a way to develop students' capacity and confidence as learners – a belief that is supported by the work of Alexander (2008). Further examples are given in Webb and Jones (2009).

At the more summative end of the dimension, there are also different degrees of formality, from the examination taken under strictly controlled conditions to the more informal exercises set and judged by the teacher. What is described as 'informal summative' in Fig. 3.2 may involve similar practice to 'formal formative'. However, the essential difference is the use made of the information. Here, it is use-

ful to consider the two kinds of feedback referred to earlier: feedback from teacher to students (affecting their learning) and feedback from students to teacher (affecting their teaching). In those cases, where summative assessment takes place at the end of a course or unit or the end of a semester or year, it may not be possible for the assessment result to be used to help these particular students improve their work. But the teachers may still use the result to adjust their approach when teaching similar activities to other students in the future (i.e. formative for teachers rather than formative for learners). So, if the results are used to adapt teaching or improving the student's learning, then it is 'formal formative'. If there is no feedback into teaching or learning, then it falls into the category of 'informal summative', even though the evidence may be the same classroom test. Ideally, of course, summative tests should be given, while there is still time for students to learn from their mistakes. When this is the case, any assessment can have a formative function, in accordance with the principle that 'Assessment of any kind should ultimately improve learning' (Harlen 2010, p. 31).

However, the process sequence represented in Fig. 3.1 should not be understood as implying that evidence to be used for summative judgements is only collected and recorded at the end of a teaching episode. Teachers may collect summative assessments en route and so build up for each student a collection of data, including pieces of students' work produced on several occasions, which will form the basis, perhaps with a formal test, for the final summative assessments. This will be expanded later in the chapter.

The Effect of Country Differences in Existing Assessment Practice

It is important to acknowledge differences among countries in respect of how assessment for formative and summative purposes is practised. Some of these differences relate to a country's traditions and existing practices. For example, whether students are assessed at the end of secondary school by external or school-based examinations has implications for classroom practice and the extent to which teachers see assessment as serving formative and summative purposes. Other differences are evident in how criteria to assess students on knowledge, skills and competencies are made explicit in the official curriculum documents and how teachers use these criteria. Black (2015) analyses eight different country practices of formative assessment and the difficulties in implementing them, often due to domination of summative assessment. As for linking formative and summative use of assessment, these variations seem relevant:

- Recognition of the need to share goals and assessment criteria with students, which may require teachers to reformulate the goals and criteria so that they can be understood by students, is a procedure that is not common practice in all countries.

- Clarity about the formative or summative status of any assessment event. In some countries it is the expectation of the students, and their parents, to be told when the student's product is to be used for summative assessment. This information affects the way students respond to a task; if a more formative purpose is intended, errors are seen as something that can help learning, while if the task is seen as for summative assessment, errors are seen as indicating that a learning goal has not been met. To put this another way, students may perform in one way if they understand that their products will be used (and judged) in order to improve their learning but may perform in a quite different way if they know the product will be judged for summative purposes. This links to the so-called didactical contract. If students know that the report has a formative purpose, they will be more experimental, more open and less tactical and will strive less for perfection, for flawless work.
- The balance between assessment for accountability purposes and assessment for learning or the dominance of a testing regime on everyday practice. National tests and the stakes of the examinations have a strong influence on teachers' teaching and students' behaviour. A survey among Danish upper secondary students revealed that 55% of the tested students have stress symptoms – due to a daily fight for high grades (Nielsen and Lagermann 2017). Their level of stress is similar to that of the 20% most stressed people in the Danish labour market (ibid.). This gives a general context for formative assessment. No matter how much teachers emphasise the formative purpose of an assessment, the summative element does not disappear. Students do seldom believe that any assessment only has a formative purpose, given today's reality in contemporary educational systems.

Approaches to Linking Formative and Summative Assessment

Given the shortcomings of tests outlined earlier, and given the wash-back effect of summative assessment on the teaching and learning, it is worthwhile looking for ways to diminish the negative effects on learning from summative use of assessment and to improve the formative use of assessment. In particular, we need ways that are capable of gathering data about the full range of competencies and understandings that are the aims of inquiry-based science, technology and mathematics education. It is, fundamentally, a question of optimising the alignment between the learning goals, the pedagogy and the assessment method (used for both formative and summative purposes). Here, linking the formative and the summative uses of assessment is a key issue. How to do this best will depend on the national and local context, but basically there are two different approaches to linking formative and summative purposes of assessment:

- *Connecting* the formative and summative use of assessment evidence. Here any assessment is used for either a formative or a summative purpose, but the methods are very alike giving relatively high validity to the summative use and relatively high reliability to the formative use.

- *Combining* the formative and summative use of assessment evidence, i.e. using evidence from an assessment for both formative and summative purposes. Using evidence from a summative assessment for formative purposes is relatively unproblematic and well-known. More innovative and full of perspective is the use of formative assessment processes for summative purposes.

Connecting the Formative and Summative Use of Assessment

The formative and summative use of assessment can be connected through:

(a) Designing a summative assessment method, an examination or a test, that is in alignment with the formative processes in everyday class. The assessment method must reflect the central goals and be able to assess the competencies in a valid way. Such an examination will replicate everyday teaching. During normal teaching students have become familiar with the criteria for fulfilling the learning goals via various formative processes. At the examination students are put in a semi-authentic situation and given a fixed time span (4 h, 8 h, 24 h, 48 h - depending on the subject) to go through (most of) the inquiry (including experimental) framework. The work is monitored and assessed (using a template with criteria known from everyday practice) and may be supplemented with an individual oral examination (both with an external examiner). Many Danish science and technology examinations have been designed and introduced following these principles, and they have the same reliability properties and the same cost level as ordinary oral examinations. Even advanced generic competences like innovation can be validly assessed for summative purpose using such authentic approaches (Nielsen 2015). Implementing an examination format in alignment with the learning goals will turn the often inexpedient wash-back effect from examination into a more productive effect.

(b) Using similar or the same assessment methods separately for both formative and summative purposes distributed throughout the year and very much performed close in time. Each assessment event is clearly identified as being used either for formative or for summative purposes. The summative assessment events are aggregated together for accountability purposes. The formative assessments are both similar to those for summative purposes but also supplemented with more valid methods able to measure more advanced competences. This is consistent with the didactical contract. Students can be confident that they will be clear about the assessment purpose – whether what they produce and present will only be formatively assessed or whether they should prepare for the summative assessments. This approach acknowledges the present accountability regime and the difficulties in changing the examinations. But it makes a clear distinction between the two purposes of assessment and thus minimises the distortion of the summative assessment.

There is considerable future potential to enhance assessment by integrating IT-based elements. The TELS (Technology-Enhanced Learning in Science) project is an early example of an IT-based system designed to improve learning and deliver assessments for summative purposes. The project has designed a Web-based Inquiry Science Environment (WISE) as an online platform for designing, developing and implementing science inquiry activities (Linn and Chiu 2011). Such new third-generation assessment systems are emerging, trying to serve both institutional (summative) purposes and individual learning (formative) purposes. They typically build a system of complex computer simulations and other performance tasks into a blended learning system. As an example, Bennett (2015) describes how Bennett and Gitomer (2009) designed an integrated system of assessments built into a demonstration programme, CBAL (Cognitively Based Assessment of, for and as Learning). The system would consist of distinct summative and formative assessment components each directed at achieving different purposes.

As Bennett (2015) writes, '… a key characteristic of third-generation assessments will be the use of complex simulations and other interactive performance tasks that replicate important features of real environments, allow more natural interaction with computers, and assess new skills in more sophisticated ways' (p. 387). Students can work with such continuously embedded assessment systems during the year, and frequent sampling of evidence of learning can then be used formatively and automatically scored and aggregated for accountability purposes. This needs to be organised in a way that does not violate students' premises for their work and their learning engagement, e.g. with a clear distinction between formative and summative use of the data and with other ethical considerations, as described by Timmis et al. (2016). The teacher's control over the system and the system's flexibility and its potential for adaptation to the teacher's own teaching approach and to the needs of the class are other important aspects of new integrated assessment systems. To some extent, there is a danger that these systems may lead to relatively rigid teaching practices. A more flexible platform was developed within the ASSIST-ME project. It was designed to support the development of students' experimental competence by providing guidance and feedback during a learning cycle. The platform is described in Chap. 9 in this book.

A tight alignment between everyday practice and the test or examination is only preferable if the test is in valid accordance with the learning goals or if the two purposes of the test are clearly separated. If this is the situation, they will minimise some of the more pernicious wash-back effects from tests and examinations. But summative assessment situations will always have some unfortunate side effects. The unusual and high-stakes-related situations will make many students nervous to a degree that affects their performance, and many of the resources invested in tests could be better used for learning activities. This is why there is a huge potential in trying to combine the formative and the summative uses of assessment evidence via extracting data for summative use from formative processes.

Combining the Formative and Summative Use of Assessment

Good assessment requires valid evidence, relevant criteria (goals) and a reliable method of judging the extent to which criteria have been met. If students are learning through inquiry, then the evidence of their learning exists in what they do in inquiry-based activities. When teachers are using assessment formatively, they are using this evidence to support learning, as in the formative assessment cycle. The activities that enable students to develop understanding and competencies are at the same time opportunities for collecting data about progress in their learning. This evidence can be accumulated from lessons on a topic and brought together at the time when achievement is to be reported and recorded.

A suggestion for how this might be done is represented in the complete Fig. 3.1, where the black arrow leading to judgement for summative use originates from the evidence gathered in the formative assessment cycle. There are several key points to make about this model.

First, formative judgements of what action, if any, to take are made in relation to the lesson or activity goals. For summative assessment judgements, however, the goals are the medium-term goals that are the targets for achievement over a period of time such as a semester or year. That is, evidence selectively accumulated over several activities is judged for summative reporting against medium-term goals stated in a national curriculum document or in a school's curriculum plan for a particular year or stage. For example, lesson or activity goals in relation to a topic on sound might be to recognise 'how sounds can be made in different ways and can be changed in loudness and pitch'. The medium-term goal, achieved after several lessons with different activity goals, might be to understand 'that sound is caused by vibration in a material'.

A second point is to emphasise that it is the evidence gathered for formative assessment that is used as evidence for summative assessment, not the formative judgements about what action to take. This is because, as noted earlier, formative assessment judgements are often not strictly criterion-referenced, and formative judgements take into account the circumstances that affect individual or groups of students. It is entirely appropriate for this to be the case in formative assessment since it means that students have the tailored help they need to take them from where they are towards the learning goals. But evidence used to report on learning for formal summative purposes has to be judged in the same way for all students, not influenced by student-related considerations.

Third, the evidence used is the 'latest and best' selected from what has been gathered, by the methods discussed earlier, over the time for which achievement is summarised and reported. The use of 'best' evidence is in recognition that over the course of a semester or year of work – at the end of which a summative report is needed – students make progress, and evidence from earlier work will be superseded by later work. 'Latest' recognises that some evidence will have been collected early in the year when particular topics were the focus of activities and may not have been revisited. There is no need to retain every piece of evidence produced during

that time, which would in any case be too burdensome. Rather, the aim is to create a collection (which could be in various forms, such as a portfolio, folder or computer file) of what is considered to be the *best* evidence of achievement. As work proceeds during the year, better work may replace earlier attempts. Involving students in making this collection helps their understanding of the quality of work that is expected and adds to the formative value of the assessment process. It also means that students understand how their work is judged for formative and summative purposes.

Fourth, formative evidence that can be used in this way has to be in a tangible form. Given the considerable diversity in the ways in which formative assessment is practised, a question arises as to whether all kinds of evidence are suitable for use in summative assessment. In the very informal methods, evidence is picked up by teachers by observation as they interact with individuals or groups of students; judgement is almost immediate, and decisions is taken quickly about any action needed. The evidence in this case is ephemeral, with no record being made, but nevertheless it provides information that is used in promoting further learning. In more formal methods, the teacher will have planned to obtain evidence, by asking students either to give written answers to questions designed to probe their knowledge and ideas or, as they conduct an inquiry, to explain their plans at intermediate stages in the work – thereby composing their own diaries of an inquiry. In between these different ways of collecting evidence are other means through which students record their work and ideas. If evidence from formative assessment is to be used for summative purposes, it follows that some, at least, of this has to be in a tangible form rather than an ephemeral such as picked up informally. This argues for using a range of ways of gathering evidence for formative assessment.

Fifth, the process of summative judgement should include measures to ensure reliability appropriate to the use of the summative assessment. At the time when a summative report is required, the best evidence is brought together and summarised by scanning for evidence of meeting the criteria indicated in the medium-term goals. The judgement can be reported in various ways. For some uses at some stages, it is only necessary to report whether goals have or have not been met. In other cases, the report may distinguish different degrees of meeting the goals, in which case there is a need to establish and exemplify the criteria for partial achievement.

Whether the judgement of accumulated best evidence is made by the teacher or by someone else depends on the intended use of the summative data. The two main purposes (leaving aside national and international monitoring for which only a sample of students are assessed) are for reporting on individual students' achievement (low stakes for teacher but high stakes for students) and recording the progress of groups of students as part of school evaluation and accountability procedures (high stakes for teachers and schools).

It is usually students' own teachers who make the judgements for regular reporting to parents and information for school records. For these audiences, scores, levels or grades give little information about what has actually been achieved. Rather, information in summary form about the extent to which goals have been achieved is

best accompanied by a narrative, identifying what has been learned well, what needs more attention and other information that will help future learning. Face-to-face interviews with parents and students can use examples of work that show what is needed to achieve certain goals.

Where the result is to be used for making decisions affecting a student's future, as in the case of some external examinations or for use in high-stakes accountability measures having another teacher or external assessor involved in the judgements, or using some moderation procedure may be considered necessary for confidence in the result. However, only if we had assessments that captured the full meaning of the learning goals, without construct over- or under-representation (see previous section on validity and reliability), and did so with optimum reliability, would it be acceptable to use students' achievement as the sole measure of school effectiveness or the only basis for accountability of schools and teachers.

Application to Inquiry-Based Education in Science

Inquiry-based science education aims to help students to develop an understanding of key ideas about science and of science (i.e. the nature of science) and ability to conduct science inquiries. Besides helping to develop these science-specific competencies, the assessment processes serve to build students' more generic competencies, such as innovation and creativity and other life and learning skills mentioned at the beginning of this chapter. The model of summative assessment based on evidence from classroom activities requires the identification of clear goals for each activity, contributing to medium-term goals for particular stages, which in turn contribute to overall long-term goals.

During the activity the teacher and students work towards the specific activity goals, gathering and using evidence from the ongoing work in cycles of formative assessment. With the medium-term goals in mind, the teacher also helps students make connections between the ideas emerging from different activities. In turn, these medium-term goals are part of the story of progress towards the overall aims of developing 'big ideas' and inquiry skill competencies. The progress towards the big ideas of science education has been mapped out in the publication *Working with Big Ideas of Science Education* (Harlen 2015).

The model of gathering data from the range of classroom activities is particularly appropriate for inquiry skills or transversal competencies such as 'investigation' or 'argumentation'. The reason for this is that the ability to apply such skills or competencies is highly dependent on the nature of the subject matter and content in which they are used, and so evidence from a range of different inquiries is needed for reliable assessment. In particular, for some inquiry topics, a clear understanding of the concepts involved may be essential, whereas for others such understanding is not needed – and indeed work to test explanations for an observed effect may help develop understanding of the relevant concept.

However, there is less clarity about progression in developing competencies relating to conducting inquiries than there is in the case of the development of sci-

entific conceptual understanding. Thus for the model to be applied, there is a need for further work to identify criteria to be used in judging the accumulated evidence relating to development of inquiry skill (described as practices in the US framework (NRC 2012)).

There is obviously a problem in that national, high-stakes, assessments which are restricted to the use of externally set tests, to be completed in writing in a strictly limited time, cannot give results which are a valid measure of inquiry competences. This problem could only be overcome if a proportion of the high-stakes results were to be based on teachers' own assessments using a variety of assessment occasions, with evidence collected over time in the form of student portfolios. This approach has been used in a few state systems, notably in Australia (Wyatt-Smith et al. 2010). The findings of that work and those from a limited attempt to develop such work with a group of teachers in England (Black et al. 2011) show (a) that it can take one or more years for teachers to develop the necessary skills, (b) that interschool collaboration is essential in order to guarantee overall comparability of results and (c) that teachers involved have found the work rewarding as it has a positive effect on their whole approach to learning and assessment.

Challenges and Benefits

The combination between formative and summative assessments may take a large variety of forms and in most of the countries raises questions. This has been discussed in the ASSIST-ME Local Working Groups in each participating country.

Most of the teachers involved in the project consider it possible to use their summative assessments (not external final exam) formatively. Using formative assessment to summative purposes raises much more divergence. In two countries (England and Slovakia), there is a large agreement, for example, teachers 'demonstrated an ability to use their formative evidence from assessment conversations to make a summative judgment in relation to inquiry competences' (cf. meeting minutes). On the other side, the law of one country (Switzerland) prohibits 'the use of information collected for formative purposes in grades'. Most of the countries, even if this practice is possible or more recommended by the official instructions, acknowledged difficulties. The main one is the 'didactic contract' in the sense that the reciprocal expectations of the teacher and the students are different depending on the way the teacher uses the students' productions to make a judgement either to help them or to give a mark. Another difficulty concerns the coherence between the criteria involved in summative and formative assessments, and with what is taught. The need of coherence is largely recognised but also the difficulty to get it. Several reasons are proposed: some competences, in particular experimental, are difficult to assess summatively with written and pencil tests, or the criteria for formative assessment are adapted according to the students, whereas in summative case, they should be the same for all. To get coherence can necessitate a long back and forth analysis between the teaching goals in terms of knowledge and competences, how they are involved in the effective teaching and their use in the different assessments.

At last the question of time is raised in all countries. The teaching constrains in terms of content do not allow the teacher to take time enough for formative assessment and to relate it to summative ones in particular in the case of learning progressions which necessitate very regular assessment on different aspects of the teaching content. Thus, if all teachers are in favour of enhancing the combination between summative and formative assessments, different ways are possible according to the culture of the countries and the practices; nevertheless all ways imply a coherence of assessment criteria. All these questions lead all the participants to suggest teacher professional development on the combination of summative and formative assessments.

Without doubt, the easiest, technical, solution to linking formative and summative use of assessment is to connect them by designing summative assessment methods able to validly and reliably assess inquiry processes. Teaching to the test would then make sense, since the test mirrors the educational goals. Such examinations have been designed (Nielsen 2015) and have been proved to fulfil standard requirements for reliability.

However, fulfilling the important requirement of validity is difficult, and school assessments in many systems seem to have a tradition which would not be justified in any other context. As Black (2016) puts it:

> the common practice of basing such assessments on a student's performance produced over only a few hours in isolation, from memory, and responding to demands that they may not have seen before, to which an immediate response in writing is required, is strange. No business enterprise would think of assessing employees' progress in this way. It is hard to justify a decision by a school, which has known about many pieces of work by a student, produced in a range of contexts and on different occasions, to base key decisions on short terminal tests.

There are several challenges to be faced in combining formative and summative use of assessment by implementing summative assessment based on evidence from classroom activities that is used for formative purpose. Among these is the need to ensure that:

- There are clear goals for classroom activities that will lead to medium-term and long-term goals.
- Teachers are assisted in translating the goals into inquiry-based activities.
- Teachers understand the role of formative assessment in learning and use a range of methods to collect relevant evidence of learning.
- Students are involved in assessing their progress and are helped to see assessment as having a positive role in learning.
- All involved in providing, collecting and using assessment data recognise that any summative assessment is necessarily an approximation.
- Judgements of teaching and the effectiveness of schools are based on a wide range of relevant evidence and not only on measures of student achievement.

However, the benefits to be gained make it essential to face such challenges. Given the weighty evidence of the value for learning – and particularly for learning through inquiry – of using formative assessment, it is important to protect its practice from overbearing and outmoded summative assessment that dominates and dic-

tates teaching and learning. Summative assessment that uses evidence from learning activities not only enables the full range of goals to be assessed but encourages, indeed requires, the practice of formative assessment.

References

Alexander, R. (2008). *Towards dialogic thinking: Rethinking classroom talk* (4th ed.). York: Dialogos.

Alonzo, A. C., & Gotwals, A. W. (Eds.). (2012). *Learning progressions in science: Current challenges and future directions*. Rotterdam: Sense Publishers.

ARG (Assessment Reform Group). (2002). *Assessment for learning: Ten principles.* http://www.hkeaa.edu.hk/DocLibrary/SBA/HKDSE/Eng_DVD/doc/Afl_principles.pdf Accessed 27 August 2016.

Baird, J.-A., & Black, P. (2013). Test theories, educational priorities and reliability of public examinations in the England. *Research Papers in Education, 28*(1), 5–21.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education, 18*(1), 5–26.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39*(1), 370–407.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York: Springer.

Black, P. (2015). Formative assessment – An optimistic but incomplete vision. *Assessment in Education: Principles, Policy and Practice., 22*(1), 161–177.

Black, P. (2016). The role of assessment in pedagogy – And why validity matters. Ch.45, pp. 725–739 in D. Wyse, L. Hayward & J. Pandya (eds.) *The Sage Handbook of curriculum, pedagogy and assessment.* London U.K.: Sage *see particularly pp. 733–734.*

Black, P., Harrison, C., Hodgen, J., Marshall, M., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education., 18*(4), 451–469.

Butler, R. (1987). Task involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest and performance. *Journal of Educational Psychology, 79*(4), 472–482.

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

Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education, 6*(1), 101–116.

Dahler-Larsen, P. (2014). Constitutive effects of performance indicators: Getting beyond unintended consequences. *Public Management Review., 16*(7), 969–986.

Dolin, J., & Krogh, L. B. (2010). The relevance and consequences of Pisa science in a Danish context. *International Journal of Science and Mathematics Education, 8*, 565–592.

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Editorial: Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching, 46*(6), 606–609.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality and development*. Philadelphia: Psychology Press.

European Commission. (2011). *Education and Training in a smart, sustainable and inclusive Europe COM* 902 final, Brussels. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0902:FIN:EN:PDF.

Harlen, W. (2010). What is quality teacher assessment? In J. Gardner et al. (Eds.), *Developing teacher assessment* (pp. 29–52). London: Open University McGraw Hill.

Harlen, W. (2012). On the relationship between assessment for formative and summative purposes. In J. Gardner (Ed.), *Assessment and learning* (pp. 87–102). London: Sage.

Harlen, W. (2013). *Assessment and inquiry-based science education: Issues in policy and practice.* Trieste: IAP. Available for free download from http://www.interacademies.net/File.aspx?id=21245. Accessed on 22 September 16.

Harlen, W. (2015). *Working with big ideas of science education*. Trieste: IAP SEP. Available for free download in English from http://www.interacademies.net/File.aspx?id=26736 and in Spanish from http://interacademies.net/File.aspx?id=28260

Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. *Assessment in Education, 20*(2), 169–207.

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

Johnson, S. (2012). *Assessing learning in the primary classroom*. London: Routledge.

Linn, M. C., & Chiu, J. L. (2011). Combining learning and assessment to improve science education. *Research and Practice in Assessment, 5*, 5–14.

McManus, S. (2008). *Attributes of effective formative assessment*. Washington, DC: Council for Chief State School Officers (CCSSO).

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement 3rd edition* (pp. 13–103). Washington, DC: America Council on Education.

Midgley, C., Kaplan, A., & Middleton, M. (2001). Performance-approach goals: Good for what, for whom, under what circumstances, and at what cost? *Journal of Educational Psychology, 93*(1), 77–86.

National Research Council (NRC). (2012). *A framework for K-12 science education* (p. 41). Washington, DC: The National Academies Press.

Nielsen, J. A. (2015). Assessment of innovation competency: A thematic analysis of upper secondary school teachers' talk. *The Journal of Educational Research, 108*, 318–330.

Nielsen, A. M. & Lagermann, L. C. (2017). *Stress i gymnasiet - Hvad der stresser gymnasielever og hvordan forebyggelse og behandling virker med 'Åben og Rolig for Unge*. DPU, Aarhus Universitet. (In Danish. Title in English: Stress in upper secondary – what stresses high school students and how to prevent and treat).

OECD. (2013). *Synergies for better learning: An international perspective on evaluation and assessment*. OECD Reviews of Evaluation and Assessment in Education, OECD Publishing: Paris. http://dx.doi.org/10.1787/9789264190658-en

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S.-y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53*, 821–846.

Pearson. (2005). *Achieving student progress with scientifically based formative assessment: A white paper from Pearson*. Referenced in Bennett 2011.

Shepherd, L. A. (2008). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279–303). New York: Erlbaum.

Smith, C., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications for children's learning for assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement, 14*(1&2), 1–98.

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

Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2016). Rethinking assessment in a digital age: Opportunities, challenges and risks. *British Educational Research Journal, 42*, 454–476.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological process*. Cambridge, MA: Harvard University Process.

Webb, M., & Jones, J. (2009). Exploring tensions in developing assessment for learning. *Assessment in Education: Principles, Policy & Practice, 16*(2), 165–184.

Wiliam, D. (2011). *Embedded formative assessment*. Bloomington: Solution Tree Press.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education., 17*(1), 59–75.