

Chapter 31

Embedding Formative Assessment in Classroom Practice

Mary James

Abstract This chapter examines the challenges of embedding formative assessment in classroom practice. It begins with a brief history of an idea that has become popular across the world in recent years. It asks why formative assessment is such a potentially powerful idea and also why it has had such uncertain impact. It looks again at what formative assessment, also known as assessment for learning, is and how practices associated with it can be improved. Sources of problems in conceptualisation and implementation are examined, and research on organisational conditions for embedding in classrooms and spreading within and across schools is described. The importance of developing both teachers' practices and their beliefs about learning is highlighted. The chapter concludes with some reflection on what might be done to put the idea back on track and considers especially the potential role of developers of tools to assist teachers and students in implementing formative assessment.

Keywords Formative assessment • AfL • Learning how to learn • Autonomous learning • Teacher learning • Student learning • Organisational learning • Implementation • Assessment tools

A Brief History of a Concept

It is usually acknowledged that Michael Scriven (1967) first proposed a 'formative/summative' distinction although he had in mind the roles performed by evaluations of educational programmes. It was Benjamin Bloom who, 2 years later, made a similar distinction with respect to students. The purpose of formative assessment, he said, was '...to provide feedback and correctives at each stage in the teaching-learning process' (Bloom 1969, p.48). The concept lay somewhat dormant for another 20 years, possibly because programme evaluation, in which Scriven was

M. James (✉)

Faculty of Education, University of Cambridge, 184 Hills Road, Cambridge CB2 8PQ, UK
e-mail: mej1002@cam.ac.uk

interested, was a dominant concern in both academic and policy circles during the 1970s.

The distinction came to prominence again in 1988, in England, when the Task Group on Assessment and Testing (TGAT) was set up by the then Conservative Government to advise on a system for assessing achievement on the national curriculum that was about to be introduced. Chaired by Professor Paul Black, this group set about defining the purposes of assessment, which they judged to be four in number: formative, diagnostic, summative and evaluative. This initiated a debate that persists to the current time and in which Paul Black has continued to be a major figure.

At around the same time, a group of UK researchers was convened by the British Educational Research Association (BERA) to provide commentary on assessment policy developments, backed by research evidence. This was known as the BERA Assessment Policy Task Group but later transformed into the UK Assessment Reform Group (ARG). (See Daugherty 2007, for an account of the work of this group.) As one of its activities, the ARG decided to seek funding to update a review, by Terry Crooks in New Zealand, of research on the impact of evaluation/assessment practices on student learning. Crooks (1988) had particularly noted the wash-back effects on student learning strategies, motivation and achievement. The ARG was successful in their bid to the Nuffield Foundation, and they asked Paul Black, and his colleague Dylan Wiliam, to carry out the new review. The result was a 35,000-word article in a refereed journal (Black and Wiliam 1998a) and a short booklet, *Inside the Black Box* (1998b). This booklet became enormously popular with teachers, teacher educators and advisers and sold tens of thousands of copies.

However, even at an early stage, there were concerns that formative assessment, as a concept, was not fully understood, so the ARG attempted to make it more transparent by distinguishing ‘assessment for learning’, as part of pedagogy, from ‘assessment of learning’ for grading and reporting. In 1999, the ARG produced another booklet, *Assessment for Learning: Beyond the Black Box*, and, in 2002, they developed a poster entitled *Assessment for Learning: 10 Principles*.

Although Paul Black held to his preference for the term ‘formative assessment’ because assessment cannot claim to be formative unless it has actually made a difference, whereas ‘assessment for learning’ can remain aspirational, the two expressions became interchangeable. Possibly for reasons that the ARG discerned, it was nevertheless ‘assessment for learning’ (AfL) that was taken up more widely, especially by policy makers. By 2008, the New Labour Government in England had introduced AfL national strategies for both primary and secondary schools backed by £150 m of government funding to provide teachers with training. The materials quoted the ARG’s definition of AfL and its ten principles. Wales, Northern Ireland and Scotland also developed AfL policies, although in Scotland, this was called Assessment is for Learning (AifL) (see James 2011, for an account of how and why these diverged). In other countries also, formative assessment or assessment for learning policies and practices developed (see James 2010 for an overview). For example, in Hong Kong, the Education Bureau’s 10-year programme of reforms, initiated in 2000, put more emphasis on assessment for learning. Even in the USA,

where psychometric approaches to measurement in education have long held sway, the reports of the Gordon Commission¹ in 2013 affirm that the primary purpose of assessment is to inform and improve teaching and learning.

With all this activity at all levels in national systems across the world, it would be reasonable to expect that teaching, learning and achievement would be transformed for the public good by innovation in formative assessment/AfL practices. Yet, in 2006 in the USA, James Popham described AfL as an ‘endangered species’ (Popham 2006). Similarly, in 2012, Dylan Wiliam was reported as saying that it was a tragedy that, despite the seeming ubiquity of AfL as an idea, in practice, the strategy is largely missing from schools in England (Stewart 2012). Indeed the term ‘assessment for learning’ has largely disappeared from the lexicon of the Department for Education, under the Conservative-led coalition government since it came to power in May 2010.

Why is it that such a potentially powerful idea, backed by evidence, has had such uncertain impact? In the next sections, I will first go back to basics to look again at what formative assessment/AfL is, before examining the sources of problems in implementation and reflecting on what might be done to put it back on track.

What Is Formative Assessment/AfL?

A central feature of all assessment is the observation of what one person says or does by another or, in the case of self-assessment, reflection on one’s own knowledge, understanding or behaviour. This is true of the whole spectrum of assessments, from formal tests and examinations to informal assessments made by teachers in their classrooms many hundred times each day. Although the form that assessments take may be very different – some may be pencil and paper tests whilst others may be based on questioning in normal classroom interactions – all assessments have some common characteristics. They all involve:

1. Making observations.
2. Interpreting the evidence.
3. Making judgements that can be used for decisions about actions.

Observation In order to carry out assessment, it is necessary to find out what students know and can do or the difficulties they are experiencing. Observation of regular classroom activity, such as listening to talk, watching students engaged in tasks or reviewing the products of their class work and homework, may provide the information needed, but on other occasions, it may be necessary to elicit the information needed in a very deliberate and specific way. A task or test might serve this purpose, but a carefully chosen oral question can also be effective. Students’ responses to tasks or questions then need to be interpreted. In other words, the assessor needs to work out what the evidence means.

¹(<http://www.gordoncommission.org/index.html>)

Interpretation Interpretations are made with reference to what is of particular interest such as specific skills, attitudes or different kinds of knowledge. These interpretations are often based on criteria that relate to learning goals or objectives. Usually observations as part of assessment are made with these criteria in mind, i.e. formulated beforehand, but sometimes teachers observe unplanned interactions or outcomes and apply criteria retrospectively. Interpretations can describe or attempt to explain behaviour, or they can infer from behaviour, e.g. what a child says, that something is going on inside a child's head, e.g. thinking. For this reason, interpretations are sometimes called inferences.

Judgement On the basis of these interpretations of evidence, judgements are made. These involve evaluations. It is at this point that the assessment process looks rather different according to the different purposes it is expected to serve and the uses to which the information will be put. This is where the formative/summative distinction becomes especially important.

In formative assessment/AfL, observations, interpretations and criteria may be similar to those employed in assessment of learning, but the nature of judgements and decisions that flow from them will be different. In essence, formative assessment/AfL focuses on what is revealed about where children are in their learning, especially the nature of, and reasons for, the strengths and weaknesses they exhibit. Formative judgements are therefore concerned with what they might do to move forward.

The Assessment Reform Group (2002) defined assessment for learning as follows:

Assessment for Learning 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.

One important element of this definition is the emphasis on students' own use of evidence. This draws attention to the fact that teachers are not the only assessors. Students can be involved in peer and self-assessment, and, even when teachers are heavily involved, students need to be actively engaged. Only learners can do the learning, so they need to act upon information and feedback if their learning is to improve. This requires them to have understanding but also the motivation and will to act. The implications for teaching and learning practices are profound and far-reaching and indicate that formative assessment should be integral to pedagogy, not an add-on.

What Does Research Say About How Formative Assessment/AfL Might Be Improved?

The generally acknowledged key source is the review of research by Paul Black and Dylan Wiliam (1998a, 1998b) mentioned earlier. In this, they analysed 250 studies of which 50 were a particular focus because they provided evidence of gains in achievement after ‘interventions’ based on what we might now call formative assessment/AfL practices. These gains, measured by pre- and post-summative tests, produced standardised effect sizes of between 0.4 and 0.7. There was evidence that gains for lower-attaining students were even greater. These findings convinced many teachers and some policy makers that formative assessment/AfL is worth taking seriously.

The innovations introduced into classroom practice involved some combination of the following:

1. Developing Classroom Talk and Questioning

Asking questions, either orally or in writing, is crucial to the process of eliciting information about the current state of a student’s understanding. However, questions phrased merely to establish whether students know the correct answers are of little value for formative purposes. Students can give right answers for the wrong reasons or wrong answers for understandable reasons. For example, Vinner (1997) showed that students gave very different answers to superficially similar questions on fractions in mathematics. When the students were asked to talk through how they had reached their answers, it emerged that many students developed a naïve conception (a rule of thumb) that large fractions have small denominators and small fractions have large denominators. This rule often serves them well, and their teachers may be unaware of the misconception. Thus, if learning is to be secure, superficially ‘correct’ answers need to be probed and misconceptions explored. In this way students’ learning needs can be diagnosed.

Research in science education, by Millar and Hames (2003), has shown how carefully designed diagnostic ‘probes’ can provide quality information of students’ understanding to inform subsequent action. The implication is that teachers need to spend time planning good diagnostic questions. Students can be trained to ask questions too and to reflect on answers. They need thinking time to do this, as they do to formulate answers that go beyond the superficial. Increasing thinking time, between asking a question and taking an answer, from the average of 0.9 of a second, can be productive in this respect. A ‘no hands up’ rule is also useful because it conveys the message that every student in the class can be called upon to answer, in the knowledge that their answer will be dealt with seriously, whether right or wrong.

All these ideas call for changes in the norms of talk in many classrooms. By promoting thoughtful and sustained dialogue, teachers can explore the knowledge

and understanding of students and build on this. The principle of ‘contingent teaching’ underpins this aspect of formative assessment/AfL.

2. Giving Appropriate Feedback

Feedback is always important, and perhaps the most powerful aspect of formative assessment practice (Hattie 2009), but it needs to be approached cautiously because research also draws attention to potential negative effects. Kluger and DeNisi (1996) reviewed 131 studies of feedback and found that, in two out of five studies, giving people feedback made their performance worse. Further investigation revealed that this happened when feedback focused on their self-esteem or self-image, as is the case when marks are given, or when praise focuses on the person rather than the learning. Praise can make students feel good, but it does not help their learning unless it is explicit about what the student has done well.

This point is powerfully reinforced by research by Butler (1988) who compared the effects of giving marks as numerical scores, comments only and marks plus comments. Students given only comments made 30% progress, and all were motivated. No gains were made by those given marks or those given marks plus comments. In both these groups, the lower achievers also lost interest. The explanation was that giving marks washed out the beneficial effects of the comments. Careful commenting works best when it stands on its own.

Another study, by Day and Cordón (1993), found that there is no need for teachers to give complete solutions when students ‘get stuck’. Indeed, students aged nine retained their learning longer when they were simply given an indication of where they should be looking for a solution (a ‘scaffolded’ response). This encouraged them to adopt a ‘mindful’ approach and active involvement, which rarely happens when teachers ‘correct’ students’ work.

3. Sharing Criteria with Learners

Research also shows how important it is that students understand what counts as success in different curriculum areas and at different stages in their development as learners. This entails sharing learning ‘intentions, expectations, objectives, goals’ and ‘success criteria’. However, because these are often framed in generalised ways, they are rarely enough on their own. Students need to see what they mean, as applied in the context of their own work, or that of others. They will not understand criteria right away, but regular discussions of concrete examples will help students develop understandings of quality. According to Sadler (1989, p. 121):

The indispensable conditions for improvement are that the student comes to hold a concept of quality roughly similar to that held by the teacher, is able to monitor continuously the quality of what is being produced during the act of production itself, and has a repertoire of

alternative moves or strategies from which to draw at any given point. In other words, students have to be able to judge the quality of what they are producing and be able to regulate what they are doing during the doing of it....

In a context where creativity is valued, as well as excellence, it is important to see criteria of quality as representing a 'horizon of possibilities' rather than a single end point. Notions of formative assessment as directed towards 'closing the gap', between present understanding and the learning aimed for, can be too restrictive if seen in this way, especially in subject areas that do not have a clear linear or hierarchical structure.

4. Peer Assessment and Self-Assessment

The formative assessment/AfL practices described above emphasise changes in the teacher's role. However, they also imply changes in what students do and how they might become more involved in assessment and in reflecting on their own learning. Indeed, questioning, giving appropriate feedback and reflecting on criteria of quality can all be rolled up in peer and self-assessment. This is what happened in a research study by Fontana and Fernandes (1994). Over a period of 20 weeks, elementary school students were progressively trained to carry out self-assessment that involved setting their own learning objectives, constructing relevant problems to test their learning, selecting appropriate tasks and carrying out self-assessments. Over the period of the experiment, the learning gains of this group were twice as big as those of a matched 'control' group.

The importance of peer and self-assessment was also illustrated by Frederiksen and White (1997) who compared learning gains of four classes taught by each of the three teachers. All the classes had an evaluation activity each fortnight. The only thing that was varied was the focus of the evaluation. Two classes focused on what they liked and disliked about the topic; the other two classes focused on 'reflective assessment', which involved students in using criteria to assess their own work and to give one another feedback. The results were remarkable. All students in the 'reflective assessment group' made more progress than students in the 'likes and dislikes group'. However, the greatest gains were for students previously assessed as having weak basic skills. This suggests that low achievement in schools may have much less to do with a lack of innate ability than with students' lack of understanding of what they are meant to be doing and what counts as quality.

From 1999 to 2001, a development and research project was carried out by Paul Black et al. (2003) at King's College London, with teachers in Oxfordshire and Medway (the King's, Medway and Oxfordshire Formative Assessment Project or KMOFAP), to test some of these findings in a British context. They found peer assessment to be an important complement to self-assessment because students learn to take on the roles of teachers and to see learning from their perspective. At the same time, they can give and take criticism and advice in a nonthreatening way

and in a language that children naturally use. Most importantly, as with self-assessment, peer assessment is a strategy for 'placing the work in the hands of the students'.

5. Thoughtful and Active Learners

The ultimate goal of formative assessment/AfL is to involve students in their own assessment so that they can reflect on where they are in their own learning, understand where they need to go next and work out what steps to take to get there. The research literature sometimes refers to this as the processes of self-monitoring and self-regulation. It could also be a description of learning how to learn. In other words, they need to understand both the desired outcomes of their learning and the processes of learning by which these outcomes are achieved, and they need to act on this understanding. Students need to become both thoughtful and active learners. They must, in the end, take responsibility for their own learning; the teacher's role is to help them towards this goal. Formative assessment/AfL is therefore, potentially, a vital tool for this purpose of promoting learning autonomy.²

Trouble with Conceptualisation and Implementation

Given all the interest in formative assessment/AfL generated in the late 1990s and claimed impact on policy and practice in the 2000s, it is perhaps surprising that success in terms of promised outcomes has remained somewhat elusive. Moreover, there has been criticism from some quarters that the advocates of formative assessment/AfL have overclaimed the benefits of a set of practices that are still not well enough conceptualised. For example, Randy Bennett (2011) identifies six areas of concern: weaknesses in the definition of formative assessment, in the basis of claims for effectiveness, in relative lack of attention to subject/domain considerations, in under-representation of measurement principles such as the validity and reliability of inferences, in underestimation of the time and support needed by teachers and in lack of attention to larger system requirements for comprehensive reform. There are reasonable grounds for some of his concerns.

In England, where assessment for learning (AfL) became enshrined in national policy for a time, understanding of the formative dimension is certainly in danger of being lost. The National Strategies of 2008 must bear some responsibility for this. They made reference to definitions of AfL and research-based accounts of good practice, but they implied that AfL can be formative, or summative, or both. The New Labour Government had invested a great deal in the development of student

²More detail of the research and advice on practical strategies can be found in James, M. et al., 2006; Wiliam, D., 2011; and Earl, L.M., 2013.

tracking and planning tools, to help teachers and principals use the results of statutory national tests for monitoring, prediction and target setting. It was politically expedient therefore to promote frequent mini-summative assessment, to secure higher performance on tests and to meet prescribed numerical targets, rather than use scarce resources on what may have appeared to be less tangible approaches to formative assessment. What was not well understood was that it is quite possible to drill students to perform well on tests without actually enhancing learning. Given the high-stakes consequences for schools that perform badly, there is increasing evidence that this is happening (Mansell et al. 2009).

Although the government in England changed in 2010, the drive is still to raise standards as measured by national curriculum tests and examinations. In fact this has intensified under the Conservative-led coalition. Nuanced ideas, about the role of formative assessment/AfL in pedagogy to enhance the learning of capable, resourceful and autonomous citizens, seem almost entirely absent. Those who are convinced by research that formative assessment is the key to improved learning and achievement have still to convince those who believe that competition, generated by the pressure of regular testing and performance tables, raises standards. The struggle between these competing positions is very evident in England at the time of writing but also reflects ideological movements globally.

These debates have almost certainly influenced the extent to which teachers have felt motivated and supported to implement innovations in classroom practice. But there are other barriers and affordances. Some of these were predictable, even in the late 1990s, because they are familiar from decades of research on educational development and innovation in schools. A more recent study, specifically related to implementation and dissemination of formative assessment/AfL values and practices, illustrates the challenges.

Lessons from the Learning How to Learn Project

Many of the successful studies that Black and Wiliam reviewed were based on small-scale experiments involving interventions often carried out by researchers. However, the success of formative assessment/AfL, more generally, depends on teachers who are required to learn new knowledge, develop new skills and reassess their roles. Therefore, teachers need to learn, as well as their students, and schools need to support them in this, which requires organisational learning. As noted above, adequate support for teachers is one of Bennett's (2011) main concerns.

The 'Learning how to learn in classrooms, schools and networks' (LHTL) development and research project (James et al. 2007) set out to investigate two key questions:

- How can formative assessment/AfL practices be developed and *embedded* in classrooms without intense outside support?

- What conditions in schools and networks support the creation and *spread* of such knowledge and practices?

The project team, from five universities, worked with 40 secondary, primary and infants schools in southern England. According to performance tables and inspection reports, most of these schools were broadly ‘average’ at the start of the project, i.e. with room for improvement.

The premise of the project was that if innovations in formative assessment/AfL were to spread ‘system-wide’, they would need to be implemented in authentic settings with much less support. Thus, we chose to provide little more than the kind of help schools might find within their local authorities (school districts) or from their own resources. We then observed what happened. We were especially interested in how the project ‘landed in schools’ and why innovation ‘took off’ in one context but not another. Our particular interest was in the conditions within and across schools that are conducive to the ‘scaling up’ and ‘rolling out’ of formative assessment/AfL practices.

As one part of our data collection, 27 lessons were filmed at the midpoint of the project to provide snapshots of classroom practice. These video recordings were placed alongside evidence from interviews with the same teachers about their beliefs about learning and their students’ comments on the lessons. These snapshots also sat within a wider picture of teachers’ practices and values distilled from survey data collected from 1,200+ teachers in 32 or our 40 schools. Three main dimensions of classroom practice (factors) emerged from the wider questionnaire evidence, which provided a useful initial framework for the study of the video evidence. These related to evidence of teachers ‘making learning explicit’, ‘promoting learning autonomy’ or pursuing a ‘performance orientation’, i.e. in contrast to a learning or mastery orientation (Dweck 2000).

What became apparent from the video material was that formative assessment practices were being handled very differently in the various lessons observed. Formative assessment/AfL strategies had been adopted, in some lessons, in ways that reflected what might be called the ‘spirit’ of AfL, showing a deep understanding of the principles underpinning the practices. In other lessons, the implementation of AfL seemed more mechanical, more the ‘letter’, focusing on surface techniques. One factor in particular seemed to differentiate one type of lesson from another: promoting learning autonomy. This was associated with the way in which that principle was illustrated in the tasks that the students undertook. An example may help to illuminate the distinction we made (see also Marshall and Drummond 2006).

Two of our video recordings were of different teachers of English, teaching classes of 13-year-olds. Ostensibly, they were both attempting to do similar things in similar contexts. In both lessons, the teachers shared the criteria with the students by giving them a model of what was needed. The students then used those criteria to assess the work of their peers.

In lesson A, students were looking at a letter they had written based on a Victorian short story; in lesson B, they were asked to consider a dramatic rendition of a nineteenth-century poem. Both had the potential to enable students to engage with

the question of what constitutes quality in a piece of work – an issue which is difficult in English and hard for students to grasp. The teacher, in lesson A, modelled the criteria by giving the students a piece of writing which was full of errors. They were asked to correct it on their own. The teacher then went through the corrections with the whole class before asking them to read through and correct the work of their peers. In lesson B, the teacher and the classroom assistant performed the poem to the class and invited the students to critique their performance. From this activity, the class as a whole, guided by the teacher, established the criteria. These criteria then governed both the students' thinking about what was needed when they acted out the poem themselves and the peer assessment of those performances.

Two crucial but subtle elements differentiate these lessons. To begin with, the scope of the task in lesson A was considerably more restricted in helping students understand what quality might look like, focusing instead on those things that were simply right and wrong. Students in lesson B, on the other hand, engaged both in technical considerations, such as clarity and accuracy, as well as the higher-order, interpretive concepts of meaning and effect. In addition, the modelling of what was required in lesson B ensured that students went beyond an imitation of that model. Each of the tasks in lesson B, including encouraging the students to create their own criteria, helped them to think for themselves about what might be needed to capture the meaning of the poem in performance. In other words, the sequence of activities guided them towards autonomous learning. The procedures alone, of lesson A, were insufficient to enable this last beneficial outcome of lesson B. The question concerning teachers' own learning is as follows: what is it that led the teacher of lesson B towards a deeper understanding and interpretation (the spirit of AfL) than the teacher of lesson A?

Analysis of our questionnaire and interview data suggested that teachers' *beliefs* about learning affect how they implement formative assessment/AfL in the classroom. Much of the roll-out of AfL in England, through the National Strategies, had focused on giving teachers *procedures* to try out in the classroom without considering what they already believe about learning in the first place. Some teachers feel more able to promote student autonomy in their classrooms than others. Underpinning lesson B, for example, was the teacher's strong conviction that her job was to make her classes less passively dependent on her and more dependent on themselves and one another. Unlike the teacher in lesson A, her beliefs about learning all centred on a move towards the greater autonomy of her students.

Teachers holding views similar to teacher B were also more likely to blame themselves for students not learning rather than the students themselves (or some barrier external to the classroom). This led them to question how they might change those activities that failed or capitalise on those tasks that went well.

In understanding these findings, we could not ignore the context in which teachers in England work. At the time of our study (2001–2005), teachers and students alike worked in a system dominated by the demands of the curriculum and examinations – as is still the case. The pressure was to cover the course or teach to the test rather than take the time to explore students' ideas and understanding. In this context, we thought it important to understanding any gap between what teachers say

they believe and what they actually do in the classroom. To this end, we coded 37 transcriptions of interviews with classroom teachers. Of 16 major coding categories, one was ‘performance orientation’ (140 passages), and another was ‘barriers to student learning’ (366 passages). When these two categories coincided, we found three subcategories: ‘pressures of curriculum coverage’, ‘pressures of national testing’ and ‘pressures of a tick-box culture’.

The tensions and dilemmas that teachers face, and their struggles to bring their practice in line with their educational values, whilst coping with pressures from outside, were a strong feature of their learning in the classroom. Some appeared content with ‘going through the motions’ of trying out new practices, but a small proportion – only about 20% however – ‘took them to heart’ and, with a strong sense of their own agency, tested and developed these ideas in their own classrooms in creative ways.

The fact that implementation of formative assessment/AfL was proving to be so difficult challenged us to find out what kinds of support within and beyond schools would allow the 20% to grow to nearer 100%. Thus, we turned our attention to analysis of school-level data. We constructed a questionnaire to be administered to staff in our project schools on two occasions, 2 years apart. This had 84 items in three sections, each relating to a dimension of interest to us: classroom assessment practice and values, teacher learning practice and values and school management and systems practices and values.

Based on factor analysis, we found marked gaps between teachers’ values and their practices that were related to *promoting learning autonomy* (practices noticeably behind values) and *performance orientation* (practices noticeably ahead of values). By the end of the project, teachers were rebalancing their assessment approaches in order to bring their practices into closer alignment with their values. Schools’ performance data indicated no negative impact of these changes on school performance, as measured by national test results, and there were some significant success stories. In some of our most successful schools, there was much higher valuing and practice of promoting learning autonomy. For example, in one school with 84% 5A*–Cs at GCSE³ in 2004 and high value-added scores, the majority of teachers consistently valued *making learning explicit* and *promoting learning autonomy* highly (above *performance orientation*), and their values-practice gaps were minimal.

³ The GCSE (General Certificate of Secondary Education) is taken by students in England, Northern Ireland and Wales when pupils are 16 years of age. In each subject, there are grades (A* to G). Students who produce an exceptional performance in the highest grade, A, are awarded an A*. Great significance is attached to these grades as schools are ranked on the percent of pupils gaining at least 5 A–C grades which must include mathematics and English language. In 2017, new grades 9–1 will be introduced to allow for finer discrimination between candidates. Grade 9 will be equivalent to A*, Grades 8 and 7 to an A, Grades 6 and 5 to a B and Grade 4 to a C, in 2017.

We also carried out multiple regression analyses to look at associations between factors on the different dimensions. We wanted to find out to what extent the variation in classroom practice might be accounted for by teachers' own learning practices and/or school management practices. Our key findings indicated that what appear to be important, at the level of the school, are:

- *A clear sense of direction*: there is communication within the school of a clear vision; there is also commitment among staff to that vision.
- *Systems of support for professional development*: teachers released to plan together; they are encouraged to experiment and to take risks with their practice along with a range of other learning opportunities.
- *The management of knowledge*: expertise is audited; schools have systems for locating the strengths of staff as a basis for managing staff expertise and building on it through support for internal and external networking.

However, the impact of these school-level factors on classroom practice, particularly those practices associated with effective formative assessment/AfL, is indirect; they are mediated by teachers' own learning practices, particularly *collaborative classroom-focused inquiry*. Thus, the key school condition for the promotion of what we termed 'learning how to learn' by students appears to be development and support of teacher learning through their inquiry into classroom experience. This might include learning from research, but also working with other teachers to plan, implement and evaluate new ideas.

Data from coordinator and head teacher interviews revealed that embedding changes in classroom practice, teachers' professional learning and school systems and practices is a process that takes time and is never entirely completed since contexts change. Embedding occurs through differing combinations of approaches and practices: working groups, standing items on meetings, school and department improvement plans, teacher 'champions' working together, informal dialogue, inviting and acting on feedback from students and networking with other schools. These differing combinations of approaches and practices reflect the fact that schools have people with different strengths, dispositions and priorities, that schools are at differing stages of development and organisational maturity and that they face differing changing contexts. Within-school and between-school differences indicated a need for differentiated approaches to continuing professional development for teachers and to school improvement plans. However, each approach or practice has both structural and cultural aspects, which interplay in complex ways. The challenge for leadership, as revealed by our data, was to create space and the climate for reflection and sharing, which includes encouraging dialogue, dissent and risk-taking. We came to view 'double loop learning' (Argyris and Schön 1978) as particularly important at school level. This involves stepping back from the familiar plan-do-review cycle to examine each stage before stepping back in to do something new. This process, at organisational level, mirrors the process of strategic and reflec-

tive inquiry for teacher learning, which in turn mirrors the process of developing students' learning autonomy, through formative assessment/AfL.

In summary, then, the LHTL project illustrated the challenges of implementation with respect to formative assessment/AfL, but it also indicated ways forward.

What Is to Be Done?

I recall a discussion in the Assessment Reform Group, around 1998, at the time when we were debating whether to introduce the distinction between assessment for learning and assessment of learning. We wondered whether what we wanted to describe had much to do with assessment at all. Were we not really striving towards a new formulation of effective pedagogy? Certainly many of the elements are now encapsulated in the principles of effective pedagogy brought together by the TLRP (James and Pollard 2012).

At the end of its deliberations, the ARG decided to keep the spotlight on assessment because of a perceived need to disrupt the widespread assumption that assessment is just another word for testing and that test scores (or grades or levels) provide enough information to enable teachers and students to know what to do next in order to improve. We wanted to reappropriate the term and restore some of the meaning conveyed by its Latin roots – that 'educational assessment' involves 'sitting beside' to 'lead out'. I suspect we were only moderately successful in this because evidence suggests that frequent mini-summative assessments are often thought to be formative. Yet only if the assessment information is actually used to help students towards deeper learning, and wider and higher achievement, can it be called formative.

As other chapters in this handbook illustrate, there is now a sophisticated understanding of the theory and practice of teaching and learning and how this can be supported in different domains and by structures and processes for teacher learning. But perhaps there is still work to be done to conceptualise the role of assessment in enhancing learning, clarifying what its particular contribution might be and ensuring that system demands for accountability do not undermine it.

There is also still much work to do to convince sceptical teachers, parents, university admissions tutors and the general public that there is real value in developing formative assessment/AfL practice. For example, in Hong Kong, where huge efforts have been made over 10 years to consult and communicate with these groups, it has proved very difficult to change established beliefs that examination results are all that matter (Fok et al. 2006). A solution has been to try to unify assessment for learning and assessment of learning through school-based assessment (SBA) and emphasise the importance of feedback from assessments for personal improvement, thus diminishing the dominance of competition. By all accounts, there is still a long way to go. Moreover, Hong Kong probably reflects the challenges in many other countries, including in the West.

If the ARG definition of AfL as '... 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', remains satisfactory, then we should perhaps pay more attention to the 'process of seeking and interpreting evidence'. If we do not get this part right then, the following processes may be seriously flawed. Bennett (2011, p. 18) argues:

...we should try our best to decrease uncertainty and bias by considering data from multiple sources, occasions, and contexts; by grounding action in a sound cognitive-domain model, ideally one that accounts for key differences among student groups; and where possible, by getting input from others as to the meaning of responses from student groups about which we are less knowledgeable.

The implication is that those with technical expertise in the field of measurement can assist in developing formative assessment tools to help teachers make valid judgements. It also suggests that we may need to reconsider the relationship between assessment *for* learning and assessment *of* learning and perhaps bring them together again, as Hong Kong has attempted to do, provided that the primary goal of enhancing learning is not undermined. The Gordon Commission in the USA seemed to have had this in mind, although it is of some concern that there were no school teachers among its 32 distinguished members. Some educators might fear that without an appropriate dialogue between tool developers and tool users, the formative purposes will be distorted or simply not implemented.

These are difficult issues and not easily resolved. Each generation will probably need to work through them afresh. But, hopefully, if a balance can be struck, dialogue maintained and the growing evidence base drawn upon, formative assessment can become embedded in classrooms and fulfil its promise.

References

- Argyris, C., & Schön, D. (1978). *Organisational learning: A theory of action perspective*. Reading: Addison Wesley.
- Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box*. Cambridge: University of Cambridge School of Education.
- Assessment Reform Group. (2002). *Assessment for learning: 10 Principles*. Cambridge: University of Cambridge Faculty of Education.
- Bennett, R. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy and Practice*, 18(1), 5–25.
- Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, 5(1), 5–75.
- Black, P. & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. King's College London, School of Education (now available from GL Assessment, London).
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for Learning: Putting it into practice*. Maidenhead: Open University Press.
- Bloom, J. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *Educational evaluation: New roles, new means. The 63rd yearbook of the National Society for the Study of Education*, part 2 (Vol. 69, pp. 26–50). Chicago: University of Chicago Press.

- 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.
- Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58(4), 438–481.
- Daugherty, R. (2007). Mediating academic research: The Assessment Reform Group experience. *Research Papers in Education*, 22(2), 139–153.
- Day, J., & Cordón, L. (1993). Static and dynamic measures of ability: An experimental comparison. *Journal of Educational Psychology*, 85, 76–82.
- Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality and development*. Philadelphia: Psychology Press.
- Earl, L. M. (2013). *Assessment as learning: Using classroom assessment to maximize student learning* (Second ed.). Thousand Oaks, CA: Corwin.
- Fok, P. K., Kennedy, K. J., Chan, K. S. J., & Yu, W. M. (2006, May). *Integrating assessment of learning and assessment for learning in Hong Kong public examinations: Rationales and realities of introducing school-based assessment*. Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment, Singapore.
- Fontana, D., & Fernandes, M. (1994). Improvements in mathematics performance as a consequence of self-assessment in Portuguese primary school pupils. *British Journal of Educational Psychology*, 64, 407–417.
- Frederiksen, J., & White, B. (1997). Reflective assessment of students' research within an inquiry-based middle school science curriculum. Paper presented at the Annual Meeting of the AERA, Chicago.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Abingdon: Routledge.
- James, M. (2010). 'Educational assessment: Overview'. In E. Baker, B. McGaw, & P. Peterson (Eds.), *International Encyclopedia of Education: Third Edition* (Vol. 3, pp. 161–171). Oxford: Elsevier.
- James, M. (2011). 'Chapter 2: Assessment for learning: Research and policy' in the (dis)United Kingdom. In R. Berry & R. Adamson (Eds.), *Assessment Reform in Education* (pp. 15–32). New York: Springer.
- James, M., & Pollard, A. (Eds.). (2012). *Principles for effective pedagogy: International responses to evidence from the UK Teaching and Learning Research Programme*. Abingdon: Routledge.
- James, M., Black, P., Carmichael, P., Conner, C., Dudley, P., Fox, A., Frost, D., Honour, L., MacBeath, J., McCormick, R., Marshall, B., Pedder, D., Procter, R., Swaffield, S., & Wiliam, D. (2006). *Learning how to learn: Tools for schools*. Abingdon: Routledge.
- James, M., McCormick, R., Black, P., Carmichael, P., Drummond, M.-J., Fox, A., MacBeath, J., Marshall, B., Pedder, D., Procter, R., Swaffield, S., Swann, J., & Wiliam, D. (2007). *Improving learning how to learn in classrooms, schools and networks*. Abingdon: Routledge.
- Kluger, A., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119, 254–284.
- Mansell, W., James, M., & the Assessment Reform Group (2009). *Assessment in schools. Fit for purpose? A commentary from the Teaching And Learning Research Programme*. London: Economic and Social Research Classroom (ESRC)/Teaching and Learning Research Programme (TLRP). Downloadable from <http://www.tlrp.org/pub/commentaries>
- Marshall, B., & Drummond, M.-J. (2006). How teachers engage with assessment for learning: Lessons from the classroom. *Research Papers in Education*, 21(2), 133–149.
- Millar, R., & Hames, V. (2003). *Towards evidence-based practice in science education 1: Using diagnostic assessment to enhance learning*. Teaching and Learning Research Programme Research Briefing No. 1. Cambridge: University of Cambridge Faculty of Education. Downloadable from http://www.tlrp.org/pub/documents/no1_miller.pdf

- Popham, W. J. (2006). All about accountability / assessment for learning: An endangered species? *Educational Leadership*, 63(5), 82–83.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119–144.
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Chicago: Rand McNally.
- Stewart, W. (2012, July 13). Think you've implemented Assessment for Learning? *TES Magazine*.
- Vinner, S. (1997). From intuition to inhibition – mathematics education and other endangered species. In E. Pehkonen (Ed.), *Proceedings of the 21st Conference of the International Group for the Psychology of Mathematics Education*, 1, (pp. 63–78). Lahti, Finland: The University of Helsinki Lahti Research and Training Centre.
- Wiliam, D. (2011). *Embedded formative assessment*. Bloomington: Solution Tree Press.