LEADERSHIP [*for*] TEACHER LEARNING

Creating a Culture Where All Teachers Improve so That All Students Succeed

DYLAN WILIAM

Hawker Brownlow EDUCATION

[*Chapter* 4]

Formative Assessment

The relationship between instruction and what is learned as a result is complex. Even when instruction is well designed and students are motivated, increases in student capabilities are, in general, impossible to predict with any certainty. Moreover, this observation does not depend on any particular view of what happens when learning takes place.

For most of the last two thousand years, the dominant view about what happens when learning takes place was that associations were made between mental states. Learning was characterised by links between particular stimuli and particular responses. Because it is impossible to predict in advance how much practice will be required before the associations are established, determining what has been learned and then taking appropriate remedial action is essential. Within this view of learning, when students fail to learn, it indicates that the links between the stimuli and the desired responses are not strong enough, and they need to be reinforced through further practice.

While this view of learning is often disparaged as old-fashioned today, it does explain some aspects of learning, such as learning multiplication facts, quite well. That said, there are many aspects of learning that such a view cannot explain. In such an "associationist" view of learning, student errors are random. However, in the second half of the 20th century, there was increasing evidence that in many areas of learning, particularly in mathematics and science, student errors were far from random.

For example, when children between the ages of four and seven are asked, "What causes the wind?" a common response is, "Trees." Now, this is not the result of misremembering what they have been taught, nor is it the result of poor-quality science

instruction. It is the result of students constructing a model of how the world works on the basis of their experiences. This idea that learning is an active, constructive process is sometimes called "constructivism". The important thing about constructivism is that it is a view about what happens when learning takes place, rather than a specific approach to teaching. It means not assuming that a student is a blank slate but accepting that the student will already have ideas about things on which, as yet, he or she has had no formal instruction. In particular, it means finding out what students already know or believe about something before trying to teach them anything else that builds on that knowledge.

Other perspectives on learning focus on the fact that what gets learned is strongly tied to the context of the learning – that learning is situated. For many years, psychologists studied this as a problem (and specifically, a failure) of transfer – in other words, how much of the learning that happens in one context can transfer to another? However in recent years, situated approaches to studying learning have looked at the features that are present in the learning environment and to which the learner becomes accustomed, so that the learner is less likely to be able to demonstrate the same capabilities when those features are not present.

The important point here is that the idea that we need to review what our students have learned regularly and frequently is independent of any particular view of what happens when learning takes place. Whatever one's beliefs about what happens when learning takes place, the empirical evidence is that students do not necessarily – or even generally – learn exactly what they are taught and that, to be effective, teachers have to find out what their students did actually learn before moving on.

A second important point about the idea that what students learn is not generally predictable is that this does not depend on what it is that students are meant to be learning. For example, if being good at history is thought of as being more knowledgeable about facts and dates, then the teacher needs to establish which facts and dates have been learned and which have not. If, on the other hand, progress in history is defined as being able to construct historical arguments with an understanding of chronology, cause and effect, and the role of evidence, then different assessments would be used. Different views about the nature of a subject will result in different ways of finding out what students have learned, but the principle that we need to find out what students have learned before moving on is applicable to any learning.

Of course, the idea that effective instruction requires frequent checks for understanding has been around for a very long time, but about fifty years ago, a number of researchers and writers involved in education began to think of this process of checking for understanding explicitly as a form of assessment. This has a downside, in that whenever people hear the word *assessment*, they think of the formal mechanisms for

determining what students know, such as quizzes, tests and examinations, rather than more immediate but less easily interpreted sources of evidence, such as student facial expressions. However, it also represents an important advance because thinking about checking for understanding explicitly as an assessment process focuses on the quality of the evidence that teachers have for the instructional decisions they need to make.

Moreover, the practice of many, if not most, teachers is relatively underdeveloped in terms of eliciting and making use of evidence about student achievement. Teachers' instructional decisions are often based on evidence that is not representative of the learning needs of the group as a whole. In that sense, developing the ways that teachers check for understanding represents some of the lowest-hanging fruit in improving the quality of instruction. As David Ausubel (1968) remarked years ago:

> If I had to reduce all of educational psychology to just one principle, I would say this: The most important single factor influencing learning is what the learner already knows. Ascertain this and teach him accordingly. (p. vi)

The remainder of this chapter looks at the development of thinking about checking for understanding as a process of assessment. As is common in much of educational research, and as we saw in chapter 2, sometimes the same ideas are described with different terms, and sometimes the same term is used to describe very different ideas, so the assumptions underlying different approaches are examined and related to the evidence about the impact on student achievement. In the final part of the chapter, I show how formative assessment can be used as a unifying framework to draw together a number of other important current ideas in education, such as differentiated instruction and response to intervention. I will also show how classroom formative assessment can be used to highlight the particular aspects of generic teacher evaluation frameworks, such as Danielson's Framework for Teaching, that will have the greatest impact on student achievement.

The Origins of Formative Assessment

It appears to be widely accepted that Michael Scriven (1967) was the first to use the term *formative* to describe evaluation processes that "have a role in the on-going improvement of the curriculum" (p. 41). He also pointed out that evaluation "may serve to enable administrators to decide whether the entire finished curriculum, refined by use of the evaluation process in its first role, represents a sufficiently significant advance on the available alternatives to justify the expense of adoption by a school system" (pp. 41–42), suggesting "the terms 'formative' and 'summative' evaluation to qualify evaluation in these roles" (p. 43).

Two years later, Benjamin Bloom (1969) applied the same distinction to classroom tests:

> Quite in contrast is the use of "formative evaluation" to provide feedback and correctives at each stage in the teaching-learning process. By formative evaluation we mean evaluation by brief tests used by teachers and students as aids in the learning process. While such tests may be [marked] and used as part of the judging and classificatory function of evaluation, we see much more effective use of formative evaluation if it is separated from the [marking] process and used primarily as an aid to teaching. (p. 48)

Benjamin Bloom and his colleagues continued to use the term *formative evaluation* in subsequent work, and the term *formative assessment* was routinely used in higher education in the United Kingdom to describe what we might call "any assessment before the big one". The term did not feature much as a focus for research or practice in the 1970s and early 1980s, and where it did, the terms *formative assessment* and *formative evaluation* generally referred to the use of formal assessment procedures, such as tests, for informing future instruction (e.g. Fuchs & Fuchs, 1986).

In a seminal paper titled "Formative Assessment and the Design of Instructional Systems", Royce Sadler (1989) argued that the term *formative assessment* should be intrinsic to, and integrated with, effective instruction:

> Formative assessment is concerned with how judgments about the quality of student responses (performances, pieces or works) can be used to shape and improve the student's competence by short-circuiting the randomness and inefficiency of trial-and-error learning. (p. 120)

He also pointed out that effective use of formative assessment could not be the sole responsibility of the teacher, but also required changes in the roles of learners:

> 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. (p. 121)

The need to broaden the conceptualisation of formative assessment beyond formal assessment procedures was also emphasised by Torrance (1993):

> Research on assessment is in need of fundamental review. I am suggesting that one aspect of such a review should focus on formative assessment,

> that it should draw on a much wider tradition of classroom interaction studies than has hitherto been acknowledged as relevant, and that it should attempt to provide a much firmer basis of evidence about the relationship of assessment to learning which can inform policy and practice over the long term. (p. 341)

It seems clear, therefore, that while the origins of the term *formative assessment* may have been in behaviourism and mastery learning for at least two decades, there has been increasing acceptance that an understanding of formative assessment as a process has to involve consideration of the respective roles of teachers and learners.

The Definition of Formative Assessment

In 1998, Paul Black and I (1998a) published a review of the research on the effects of classroom formative assessment, which was intended to update the earlier reviews undertaken by Natriello (1987) and Crooks (1988), mentioned in the previous chapter. To make the ideas in their review more accessible, we produced a paper for teachers and policy makers that drew out the implications of their findings for policy and practice (Black & Wiliam, 1998b). In this paper, we defined formative assessment as follows:

> We use the general term assessment to refer to all those activities undertaken by teachers – and by their students in assessing themselves – that provide information to be used as feedback to modify teaching and learning activities. Such assessment becomes formative assessment when the evidence is actually used to adapt the teaching to meet student needs. (p. 140)

Some authors have sought to restrict the meaning of the term to situations where the changes to the instruction are relatively immediate:

- "The process used by teachers and students to recognise and respond to student learning in order to enhance that learning, during the learning" (Cowie & Bell, 1999, p. 32).
- "Assessment carried out during the instructional process for the purpose of improving teaching or learning" (Shepard et al., 2005, p. 275).
- "Formative assessment refers to frequent, interactive assessments of students' progress and understanding to identify learning needs and adjust teaching appropriately" (Looney, 2005, p. 21).
- "A formative assessment is a tool that teachers use to measure student grasp of specific topics and skills they are teaching. It's a 'midstream'

> tool to identify specific student misconceptions and mistakes while the material is being taught" (Kahl, 2005, p. 11).

The Assessment Reform Group – a group of scholars based in the United Kingdom and dedicated to ensuring that assessment policy and practice are informed by research evidence – acknowledged the power that assessment had to influence learning, both for good and for ill, and proposed seven precepts that summarised the characteristics of assessment that promotes learning (Broadfoot et al., 1999, p. 7):

1. It is embedded in a view of teaching and learning of which it is an essential part.
2. It involves sharing learning goals with pupils.
3. It aims to help pupils to know and to recognise the standards they are aiming for.
4. It involves pupils in self-assessment.
5. It provides feedback which leads to pupils recognising their next steps and how to take them.
6. It is underpinned by confidence that every student can improve.
7. It involves both teacher and pupils reviewing and reflecting on assessment data.

In looking for a term to describe such assessments, the group suggested that because of the variety of ways in which it was used, the term *formative assessment* was no longer helpful:

> The term "formative" itself is open to a variety of interpretations and often means no more than that assessment is carried out frequently and is planned at the same time as teaching. Such assessment does not necessarily have all the characteristics just identified as helping learning. It may be formative in helping the teacher to identify areas where more explanation or practice is needed. But for the pupils, the marks or remarks on their work may tell them about their success or failure but not about how to make progress towards further learning. (Broadfoot et al., 1999, p. 7)

Instead, they preferred the term *assessment for learning*, which they defined 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" (Broadfoot et al., 2002, pp. 2–3).

The earliest use of the term *assessment for learning* appears to be as the title of a chapter by Harry Black (1986). It was also the title of a paper given at AERA in 1992

(James, 1992), and three years later, was the title of a book by Ruth Sutton (1995). The origin of the term is often mistakenly attributed to Rick Stiggins as a result of his popularisation of the term (e.g. Stiggins, 2005), although Stiggins himself has always attributed the term to other authors.

Most recently, an international conference on assessment for learning in Dunedin, New Zealand, in 2009, building on work done at two earlier conferences in the United Kingdom (2001) and the United States (2005), adopted the following definition:

> Assessment for Learning is part of everyday practice by students, teachers and peers that seeks, reflects upon and responds to information from dialogue, demonstration and observation in ways that enhance ongoing learning. (Klenowski, 2009, p. 264)

The phrase *assessment for learning* has an undoubted appeal, especially when contrasted with assessment *of* learning, but as Bennett (2011) points out, replacing one term with another serves merely to move the definitional burden.

More important, as Paul Black and I and our colleagues have pointed out, the distinctions between assessment for learning and assessment of learning on the one hand, and between formative and summative assessment on the other, are different in kind. The former distinction relates to the purpose for which the assessment is carried out, while the second relates to the function it actually serves. We clarified the relationship between assessment for learning and formative assessment as follows:

> Assessment for learning is any assessment for which the first priority in its design and practice is to serve the purpose of promoting students' learning. It thus differs from assessment designed primarily to serve the purposes of accountability, or of ranking, or of certifying competence. An assessment activity can help learning if it provides information that teachers and their students can use as feedback in assessing themselves and one another and in modifying the teaching and learning activities in which they are engaged. Such assessment becomes "formative assessment" when the evidence is actually used to adapt the teaching work to meet learning needs. (Black, Harrison, Lee, Marshall & Wiliam, 2004, p. 10)

Arguing about different definitions of formative assessment and assessment for learning may seem like the most self-indulgent of academic debates, but definitions are important. We cannot begin to assemble evidence about the impact of formative assessment on student achievement until we establish what, exactly, it is, and the range of practices that are described as formative assessment is indeed great. At one extreme, Leahy, Lyon, Thompson and Wiliam (2005) emphasised formative

assessment as an inherent part of effective instruction, with teachers using evidence from assessment to make instructional adjustments minute by minute and day by day. At the other extreme, many commercial test publishers have produced large collections of test items (often called "formative assessment item banks") that can be used to construct tests to gauge students' progress toward valued goals (Educational Testing Service, 2010). The idea is that these tests can be keyed to the curriculum pacing guide being used in a school system to help administrators determine whether students are on track to pass state-mandated tests. These assessments are often called "interim" tests, or "benchmark" tests, although there is no universally agreed-upon definition of these terms.

Shepard (2008) suggests that the term *formative assessment* should not be applied to the interim and benchmark assessments – not out of any desire to control the language used, but because "the official definition of formative assessment should be the one that best fits the research base from which its claims of effectiveness are derived" (p. 280). She points out that the research literature that is usually invoked to justify the use of benchmark and interim assessments actually does not support their use, because the theory of action for benchmark and interim assessments is quite different from the research evidence that does exist about the effectiveness of formative assessment that is more closely tied to the instruction.

While I can understand Shepard's view, my belief is that, not least because of the business interests at stake, it is unlikely that it will be possible to restrict the use of the term *formative assessment* to what the research evidence actually shows has an impact on learning. For this reason, I believe that it makes sense to define the term *formative assessment* broadly and then to examine which kinds of approaches to formative assessment are likely to have the greatest impact on student achievement.

The following definition, proposed by Paul Black and me, is intended to be consistent with the earlier definitions but also to be inclusive of the different approaches to formative assessment discussed so far.

> Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited. (Black & Wiliam, 2009, p. 9)

There are several implications of this definition that may be helpful to draw out explicitly.

Formative assessment is not a thing. The distinction between summative and formative is grounded in the function that the evidence elicited by the assessment actually serves, and not on the kind of assessment that generates the evidence. From such

a perspective, to describe a particular assessment as formative is to make what Ryle (1949) described as a category mistake – assigning something a property it cannot have.

As Cronbach (1971) observed, an assessment is really just a procedure for making inferences. We get students to engage in particular activities, and these generate evidence that we then interpret in order to draw conclusions. Where those conclusions are related to the student's current level of achievement, or to his or her future performance, then the assessment is serving a *summative* function. Where the conclusions are related to the kinds of instructional activities that are likely to maximise future learning, then the assessment is functioning *formatively*. The summative-formative distinction is therefore a distinction in the kinds of inferences that are supported by the evidence elicited by the assessment rather than the kinds of assessments themselves. Of course, the same assessment evidence may support both kinds of inferences, but in general, assessments that are designed to support summative inferences (i.e. inferences about current or future levels of achievement) are not particularly well suited to supporting formative inferences (i.e. inferences about instructional next steps). In other words, it is in general easier to say where a student is in his or her learning than what should be done next. It might be assumed that assessment designed primarily to serve a formative function would require, as a prerequisite, a detailed specification of the current level of achievement, but this does not necessarily hold. It is entirely possible that the assessment might identify a range of possible current states of achievement that nevertheless indicate a single course of future action – we might not know where the student is, but we know what he or she needs to do next.

Anyone – teacher, learner or peer – can be the agent of formative assessment. While much of the early work in formative assessment focused on the way that teachers could collect evidence to inform their instructional decisions, it is clear that making full use of formative assessment requires changes in the roles of learners and of their peers. In particular, teachers who have developed their own practice of formative assessment routinely report that students have taken greater responsibility for their own learning (Black & Wiliam, 2012).

The focus of the definition is on decisions rather than data. Many of those who espouse the adoption of formative assessment talk about data-driven decision making. The idea that decisions should be driven by data rather than hunch, prejudice or guesswork is, to be sure, very attractive. However, when the focus is on data, there is a tendency to collect data without any clear idea of how those data will be used. Howard Wainer (2011) points out that "data only become evidence when they are used to support a claim" (p. 148) – otherwise, they are just data. The danger with data-driven decision making, then, is that data are collected simply because they might be useful in the future. This is a particular issue with benchmark or interim

assessments because often, by the time the results of such assessments arrive, the teacher has moved on to new material. In the definition proposed earlier, however, since the focus is on decisions, rather than data, then data should be collected only after it has been determined how and when the data will be used. A teacher might check ten minutes before the end of a lesson to see whether the class understands so she knows whether she should go on or review the material already covered. As another example, a benchmark assessment system was designed so that time to use the results from benchmark assessments to plan and implement instructional changes was built into the program (Bulkley, Christman, Goertz & Lawrence, 2010). The important feature of both of these cases is that how to instructionally use the data was determined *before the data were collected* – not so much data-driven decision making as *decision-driven data collection*.

The definition does not require that the assessment is effective. Many definitions of formative assessment or assessment for learning require that to be formative, the assessment must improve the learning of students beyond what would have occurred without the assessment. However, this is a rather demanding standard to satisfy. Indeed, given the complexity of learning, it seems highly unlikely that *any* process could be guaranteed to improve the learning of all the students in a diverse group. For this reason, it is important that the definition is probabilistic rather than deterministic. All that is required for an assessment to function formatively is that the evidence elicited by the assessment is *likely* to improve the achievement of students.

The definition does not require that instruction is in fact changed. In *Embedded Formative Assessment* (Wiliam, 2011), I gave an example of a situation where a teacher asked a group of students in an Advanced Placement class to sketch the graph of $y = 1/(1+x^2)$ on dry-erase boards and then hold up their responses. The teacher could see that all the students had produced an appropriate sketch, and so she moved on. The important point here is that she did not change her instructional decision. Moving on is what she had planned to do, but now, because of the evidence she had collected, she knew that moving on was the right thing to do. So, the decision was not a better decision, because it was the decision she had planned to make. But it was a better *founded* decision because now the decision was based on evidence that moving on was indeed the right thing to do.

Another way of thinking about formative assessment that follows on from the preceding definition is that formative assessment is concerned with "the creation of, and capitalisation upon, 'moments of contingency' in instruction for the purpose of the regulation of learning processes" (Black & Wiliam, 2009, p. 6).

Before discussing this idea in detail, it may be helpful to reflect on the word *regulation* used in the previous sentence. In English, the word *regulation* has two distinct

senses: a rule that has to be followed and the idea of keeping a system performing in the way it should. It is the latter sense that is intended here. The main idea here is that instruction should be designed so that if the learning processes in which students are participating are not yielding the intended learning, then this becomes apparent, so that something can be done to put the learning back on track. One way in which this can be done is by designing moments of contingency, such as checks for understanding of instruction.

Of course, these moments of contingency do not occur in a vacuum. The way in which teachers, peers and the learners themselves create, and capitalise, on these moments of contingency involves consideration of instructional design, curriculum, pedagogy, psychology and epistemology. However, the focus on these moments of contingency in learning does restrict the focus to only those aspects of instruction that reasonably could be regarded as "assessment" and thus prevents the concept of formative assessment from expanding to subsume all of learning, thereby losing any useful focus.

Elsewhere (Wiliam, 2007), I have pointed out that moments of contingency can be synchronous or asynchronous. Synchronous moments include teachers' real-time adjustments during teaching or the way a teacher, after a class poll of student responses, suggests that students discuss their responses with a neighbour (Crouch & Mazur, 2001). Asynchronous examples include those situations where teachers get students to provide feedback for each other using a protocol such as "two stars and a wish" (Wiliam, 2011) or the use of evidence derived from student work (e.g. homework, student summaries made at the end of a lesson) to plan a subsequent lesson. Most commonly, the evidence would be used to modify the instruction of those from whom the evidence was collected. However, evidence about difficulties experienced by one group used to modify instruction for another group of students at some point in the future would also qualify, although there would, of course, be an inferential leap as to whether the difficulties experienced by one group would be relevant to a different group of students.

As Allal (1988) has pointed out, the regulation can be proactive, interactive or retroactive. Proactive regulation of learning can be achieved, for example, through the establishment of didactical situations (Brousseau, 1997), where the teacher "does not intervene in person, but puts in place a 'metacognitive culture,' mutual forms of teaching and the organisation of regulation of learning processes run by technologies or incorporated into classroom organisation and management" (Perrenoud, 1998, p. 100). For example, if a mathematics teacher creates a culture in her classroom where students are routinely encouraged to reflect on the reasonableness of their answers, then the students may be able to detect errors themselves. Similarly, where students

are encouraged to share their thinking with their peers, the chances that student learning proceeds effectively are increased.

Such didactical situations can also be planned by the teacher as specific points in time when she will evaluate the extent to which students have reached the intended understanding of the subject matter – for example, through the use of hinge-point questions (Wiliam, 2011) as specific parts of the lesson plan. While the planning of such questions takes place before the lesson, the teacher does not know how she will proceed in the lesson until she sees the responses made by students, so this would be an example of interactive regulation, in which teachers use formative assessment in real time to make adjustments to their instruction during the act of instruction. The preceding examples in which teachers reflect on instructional sequences after they have been completed, for the benefit of the particular students concerned or others, would be examples of retroactive regulation of learning.

Formative Assessment and Self-Regulated Learning

Many authors have focused on formative assessment largely as a process in which teachers administer assessments to students to ensure that the intended learning has taken place (e.g. Ainsworth & Viegut, 2006), but it is clear that for at least a quarter of a century, some authors have regarded the role of the learner as central. I have suggested (Wiliam, 1999a, 1999b, 2000) that formative assessment consisted of teacher questioning, feedback and the learner's role (essentially, understanding criteria for success, peer assessment and self-assessment), and a number of other authors have proposed similar ways of understanding formative assessment or assessment for learning. For example, Stiggins, Arter, Chappuis and Chappuis (2004) proposed that assessment for learning consists of seven strategies:

1. Provide students with a clear and understandable vision of the learning target.
2. Use examples and models of strong and weak work.
3. Offer regular descriptive feedback.
4. Teach students to self-assess and set goals.
5. Design lessons to focus on one learning target or aspect of quality at a time.
6. Teach students focused revision.
7. Engage students in self-reflection and let them keep track of and share their learning.

While it could be argued that strategies 5 and 6 are not solely focused on assessment, it seems clear that a number of authors (e.g. Bailey & Heritage, 2008; Brookhart, 2007; Popham, 2008) writing about formative assessment have been addressing the same conceptual territory, although dividing it up in different ways.

Of course, where formative assessment is presented as a number of strategies, it is not clear whether the list is in any sense complete. To address this, Leahy et al. (2005) proposed that formative assessment could be conceptualised as five key strategies, resulting from crossing three processes (where the learner is going, where the learner is right now and how to get there) with three kinds of agents in the classroom (teacher, peer, learner), as shown in Figure 4.1 (Leahy et al., 2005). This model could be criticised on the grounds that the strategies are not solely concerned with the assessment process. That said, provided that the two strategies that involve learners and peers (activating students as learning resources for one another and activating students as owners of their own learning) are interpreted as specifically focusing on moments of contingency in the regulation of learning processes, then the framework provided in Figure 4.1 does, I believe, provide a useful conceptual basis for formative assessment. More detailed explanations of the derivation of the model in Figure 4.1 can be found in Wiliam (2007) and Wiliam (2011).

<table>
<tr><th></th><th>Where the learner is going</th><th>Where the learner is now</th><th>How to get there</th></tr>
<tr><td>Teacher</td><td rowspan="3">Clarifying, sharing and understanding learning intentions and success criteria</td><td>Engineering effective discussions, tasks and activities that elicit evidence of learning</td><td>Providing feedback that moves learning forward</td></tr>
<tr><td>Peer</td><td colspan="2">Activating students as learning resources for one another</td></tr>
<tr><td>Learner</td><td colspan="2">Activating students as owners of their own learning</td></tr>
</table>

Figure 4.1. Five key strategies of formative assessment

The important point here is that within this view, self-regulated learning is an essential component of formative assessment.

The Evidence for Formative Assessment

At least in terms of the five strategies previously outlined, formative assessment is more of a framework for drawing together a number of related aspects of instruction than a single aspect of instruction. It is based on the idea that instruction presents teachers and learners with a constant stream of decisions, and the assumption being

made is that these decisions will support learning more effectively if they are based on evidence.

The research evidence that formative assessment does have a significant impact on student achievement is reviewed in detail in my earlier book, *Embedded Formative Assessment* (Wiliam, 2011). For those interested in looking in detail at the academic research in this area, a summary of the main reviews of this research can be found in Wiliam and Leahy (2015). These reviews provide solid evidence that for learners of different ages, in different countries, and for different school subjects, attention to the five strategies of formative assessment has a substantial impact on student achievement. More important, as Paul Black and I noted in our 1998 review of the research:

> Furthermore, despite the existence of some marginal and even negative results, the range of conditions and contexts under which studies have shown that gains can be achieved must indicate that the principles that underlie achievement of substantial improvements in learning are robust. Significant gains can be achieved by many different routes, and initiatives here are not likely to fail through neglect of delicate and subtle features. (Black & Wiliam, 1998a, pp. 61–62)

In addition, it is worth noting that a best-evidence review of research on improving achievement in F–12 education identified the three most cost-effective strategies for improving learning in schools as feedback, peer tutoring, and metacognition and self-evaluation (Education Endowment Foundation, 2013). In other words, the three most cost-effective educational interventions were aspects of three of the five strategies of formative assessment identified previously.

Moreover, the other two strategies of formative assessment are necessary precursors to these three strategies. Feedback cannot be given unless evidence about what is going well and what is going not as well has been collected, so the second strategy – eliciting evidence – is a necessary precursor to effective feedback. And of course, without a clear idea about the intended learning, we do not know what evidence is worth collecting. The five strategies for formative assessment presented in Figure 4.1 do therefore appear to represent a minimum set of the most high-impact approaches to improving learning.

While the evidence in favour of formative assessment is strong, a number of critiques of the research on formative assessment have appeared in recent years. Perhaps the most important is that offered by Randy Bennett (2011), who identifies six issues with the research on formative assessment. Because Bennett's is the most comprehensive and rigorous of the critiques that have appeared, it seems worthwhile to discuss each of the issues in some detail.

The Definitional Issue

Bennett rightly shows that (as previously noted) there is no agreed-upon definition of what, exactly, is meant by formative assessment, which makes any discussion difficult (although he does acknowledge that the five strategies of formative assessment discussed above do represent an attempt to operationalise formative assessment in a way that could be implemented in practice). The advantage of a comprehensive definition of formative assessment, such as that previously presented, is that rather than debating which of the approaches to defining formative assessment is "correct", attention can be given to understanding the variables along which different approaches to formative assessment can be located.

Perhaps the most important variable is the theory of action implied in the approach to formative assessment. Put simply, what exactly is the formative assessment meant to *form*? Obviously, by definition, all approaches to formative assessment emphasise the role of assessment in forming student learning, but there are important differences in the mechanism by which this improvement takes place. As previously noted, for some, the defining feature of formative assessment is that it improves learning – a focus on outcomes. Others have emphasised improvements in the quality of instruction, specifically in terms of a match between the instruction and the specific learning needs of the students being taught. And, as already noted, the focus can be on improving the quality of decisions that are made during instruction.

The mechanism of improvement is also related to this decision making. Some approaches emphasise the instruments used, such as tests, quizzes or probes, while others focus on the outcomes of those assessments, and still others emphasise the functions that evidence elicited by the assessments actually serve. I argued earlier that assuming any other than the last of these leads to contradiction, although in the future, it may be possible to show that other approaches to defining the formative-summative distinction could be coherent.

Another important difference in the various approaches to formative assessment is the role of the respective agents involved in the process and, specifically, whether students from whom evidence was elicited have to be beneficiaries of the process and whether they have to be involved. For example, if a year-seven maths teacher learns something in teaching one section in period 1 and uses this to modify her instruction of the same material for the period 2 class, would this be an example of formative assessment? Furthermore, if the teacher changes her instruction based on the outcomes of a formative-assessment process, does not involve the students, and simply teaches the next lesson more effectively, would that be formative or not?

Finally, there is the length of the formative assessment cycle of evidence, inference and action (Wiliam & Black, 1996). As noted above, some approaches to formative

assessment involve cycles of several weeks, other approaches involve week-to-week cycles, while still others focus on minute-by-minute and day-by-day classroom interaction. To make clear that all of these are potentially formative, Wiliam and Thompson (2008) suggested using the terms *short-cycle*, *medium-cycle* and *long-cycle formative assessment*, as shown in Table 4.1.

Table 4.1. Cycle Length, Focus and Impact for Different Approaches to Formative Assessment

Type	Length	Focus	Impact
Long-cycle	Four weeks to one year	Across marking periods, terms, semesters, years	Improved student monitoring and curriculum alignment
Medium-cycle	One to four weeks	Within and between teaching units	Improved student-involved assessment and teacher cognition about learning
Short-cycle	Minute to minute and day to day	Within and between lessons	Increased student engagement and improved teacher responsiveness

The Effectiveness Issue

The second issue identified by Bennett (2011) is with the evidence of effectiveness. In our review of the evidence published in 1998, Paul Black and I had explicitly rejected the idea of a meta-analysis, for reasons that should be clear after the discussion of meta-analysis in chapter 3. Here's what we wrote:

> It might be seen desirable, and indeed might be anticipated as conventional, for a review of this type to attempt a meta-analysis of the quantitative studies that have been reported. The fact that this hardly seems possible prompts a reflection on this field of research. Several studies which are based on meta-analyses have provided useful material for this review. However, these have been focussed on rather narrow aspects of formative work, for example the frequency of questioning. The value of their generalisations is also in question because key aspects of the various studies that they synthesise, for example the quality of the questions being provided at the different frequencies, is ignored because most of the researchers provide no evidence about these aspects.

> Individual quantitative studies which look at formative assessment as a whole do exist, and some have been discussed above, although the number with adequate and comparable quantitative rigour would be of the order of 20 at most. However, whilst these are rigorous within their own frameworks and purposes, and whilst they show some coherence and reinforcement in relation to the learning gains associated with classroom assessment initiatives, the underlying differences between the studies are such that any amalgamations of their results would have little meaning. (Black & Wiliam, 1998a, pp. 52–53)

However, we did indicate in a subsequent, less technical paper that the evidence we had reviewed indicated that formative assessment could be expected to have effect sizes in the range of 0.4 to 0.7. This was done to give policy makers and educators some indication of the kinds of increases in educational achievement that might be possible with formative assessment but in retrospect may have been a mistake, because, as Bennett (2011) points out, this claim has become "the educational equivalent of urban legend" (p. 12).

To be sure, the estimate of 0.4 to 0.7 is consistent with a range of other estimates available at the time and that have appeared subsequently. Meta-analyses of the effectiveness of feedback by Kluger and DeNisi (1996) and by Nyquist (2003) found that feedback improved achievement by around 0.4 standard deviations, and a more recent review of research on feedback by Shute (2008) found effect sizes of between 0.4 and 0.8. Hattie's estimate of the impact of feedback on student achievement is 0.73 (Hattie, 2008).

However, from the discussion of meta-analyses in chapter 3, it should be clear that effect sizes are likely to vary considerably in magnitude from one context to another, and it is this that many critics of the research on formative assessment appear not to appreciate. A study of twenty-four middle and high school maths and science teachers developing their practice of formative assessment over the course of a year, with student achievement measured by externally scored standardised tests, found an increase in student achievement, compared with other teachers in the same schools, of 0.32 standard deviations (Wiliam, Lee, Harrison & Black, 2004).

Some, following Cohen, have labelled this effect as "small". Hattie (2008) has pointed out that, as we saw in Figure 3.1, the average annual increase in achievement is around 0.2 to 0.4, so that effect sizes in this range should be regarded as average progress and that we would need an effect size larger than 0.4 to treat an intervention as promising. However, it is not clear whether Hattie regards the 0.4 as the annual gain we should expect or whether this is the *additional* gain, beyond the normal average annual progress, that we should want before regarding an intervention as

worth recommending (support for both interpretations can be found in chapter 2 of *Visible Learning*). The effect size of 0.32 found by Wiliam et al. (2004) might indeed be regarded as modest if it was the total progress made by the students in the study. However, in this study, this was the *difference* in progress made by students taught by participating teachers and those taught by non-participating teachers over the course of a whole year. Moreover, given that the average age of the students participating in the study was 13.5 years, by reference to Figure 3.1, it can be seen that this is a rather substantial additional impact on learning. In fact, even taking into account that the tests used in the study were closely aligned to the curriculum, we estimated that the effect size of 0.32 we found would equate to an increase in the rate of learning of approximately 70 per cent.

By any measure, this is a large impact on student achievement. A 70 per cent increase in the rate of learning would, if replicated across an entire education system, result in students achieving the standards currently achieved by year twelves by the end of year seven. If the effect of formative assessment on student achievement is only 0.2, as estimated by Kingston and Nash (2011, 2015), then this would result in students reaching the achievement of current year twelves by the end of year eight. Even if 30 per cent of the extra learning each year is forgotten by the beginning of the next year, an additional 0.2 standard deviations each year would result in US students scoring 112 points higher on PISA (eight years of education making an extra 0.14 standard deviations' progress each year is 1.12 standard deviations, or 112 points). This would take the average score of US fifteen-year-olds on PISA from 492 to 604, well ahead of students in Shanghai, who averaged 587 in 2012.

The dollar value of such increases in achievement would be extraordinary. Just in terms of the costs of education, an effect size of 0.2, if it were applied across F–12 education, would be worth $150 billion each year, and the cumulative impact on the US economy would be even larger. According to one estimate (Hanushek & Wößmann, 2010), over the next fifty years, a net increase of achievement of 0.14 standard deviations each year would have a current net value of $200 trillion.

The Domain Dependency Issue

In addressing the domain dependency issue, Bennett (2011) argues that "to be maximally effective, formative assessment requires the interaction of general principles, strategies and techniques *with* reasonably deep cognitive-domain understanding" (p. 15). At one level, this is obvious. It is inconceivable that someone armed with a range of formative-assessment strategies and techniques could teach effectively a subject he or she knew nothing about. And, at the other extreme, it is well known that those who know the subject well, but have no idea about how to teach it, are not very effective (e.g. Baumert et al., 2009).

Moreover, it is now well established that the knowledge needed to teach effectively is not routinely acquired through advanced courses in the particular subject. Teachers need what has been called *pedagogical content knowledge* by Shulman (1986) and subject-specific *knowledge for teaching* by Hill, Rowan and Ball (2005). This includes knowing what kinds of difficulties students are likely to encounter in learning a particular subject and the kinds of questions that are most likely to elicit relevant evidence. What makes a good question in mathematics tells us nothing about what makes a good question in humanities and social sciences.

However, there is a much deeper point to be made here, and that is that particular pedagogical strategies and techniques are given more emphasis in some subjects than others. For example, redrafting a piece of work after having received feedback is a staple in English but is less relevant in mathematics (at least the way mathematics is taught in most US schools). Furthermore, even when a technique is equally important in different domains, it may take a different form from one subject to another. This requires acknowledging that there are important differences between school subjects that mean that formative assessment will take different forms and will emphasise different aspects of practice in different school subjects.

This realisation that to be maximally effective teacher professional development must address the particularities of the subject being taught has resulted, in many districts, in different programs of professional development for teachers in different subjects. English teachers might be working on developing their use of the Protocol for Language Arts Teaching Observations (Grossman et al., 2009), while science teachers might be working on argumentation processes in science (Gabrielsen, 2014), and mathematics teachers might be developing their understanding of mathematical investigations (Boaler, 2009). While these approaches are effective within their respective subjects, they tend to lead to a balkanisation of school improvement efforts, with little coherence to efforts across different subjects. It also suggests to students that there is little commonality to the process of learning in different subjects.

In order to counter this, to be most effective, approaches to school improvement need to identify as many commonalities as possible across different subject disciplines, while also paying attention to the particularities of each discipline.

In the earliest work that my colleagues and I conducted on formative assessment with mathematics and science (and later, English) teachers, the professional development that we provided consisted of both generic and subject-specific components. Teachers met in generic groups, but during each one-day meeting, time was also reserved for teachers to meet in subject groups to explore aspects of formative assessment that were specific to their discipline (Black, Harrison, Lee, Marshall & Wiliam, 2003). Subsequently, as well as producing generic books for teachers, we produced

a series of subject-specific guides that explore how to maximise the impact of formative assessment in English (Marshall & Wiliam, 2006), mathematics (Hodgen & Wiliam, 2006), science (Black & Harrison, 2002) and modern foreign languages (Jones & Wiliam, 2007).

The advantage of conceptualising formative assessment as both domain specific *and* generic at the same time is that a school can ensure that practice is faithful to, and effective within, a particular discipline, while bringing greater coherence to the experiences of students. Learning intentions and success criteria may differ in character from subject to subject, but the idea of learning intentions and success criteria is equally relevant to all learning (although the learning intentions and success criteria may differ in their specificity). In some subjects, teachers will elicit evidence of achievement by questioning, while in others, they may do so by observing students engaged in complex tasks, but the important point is that there is always an intentional approach to eliciting evidence. As previously noted, feedback may prompt redrafting, extension or a completely different activity but, again, the focus is on moving learning forward. And of course, efforts to activate students as learning resources for one another and as owners of their own learning will likewise take different forms in different subjects. However, by drawing attention to these processes as universally relevant to learning, students will be able to make connections across their different experiences. This will not, of course, suddenly result in students who are able to apply their knowledge in any domain, but it is likely to make the students more able to relate learning in one area to their experiences in another.

The Measurement Issue

One of the criticisms frequently made of the term *formative assessment* is that it is just good teaching, and using the assessment label is just confusing because, as noted above, when teachers hear the term *assessment*, they immediately think of the formal mechanisms of tests, quizzes and examinations. However, as also previously noted, the important point about thinking of checking for understanding as an assessment process is that it focuses teachers on the quality of the evidence they have elicited and what conclusions can reasonably be drawn from that evidence. In his review, Bennett (2011) argues that formative assessment, as a process, gives too little attention "to the fundamental principles surrounding the connection of evidence – or what we observe – to the interpretations we make of it" (p. 16). This may, of course, be true for many approaches to formative assessment, but it is strange that he does not realise that two of the modules of the Keeping Learning on Track program developed by his own organisation are specifically focused on generating items that allow teachers to draw appropriate inferences about student achievement. This idea of developing high-quality items is also given significant attention in *Embedded Formative Assessment* and our

more recent work (Wiliam & Leahy, 2015) and is also the focus of earlier theoretical work (Wylie & Wiliam, 2006, 2007).

The last two issues identified by Bennett are those of professional development and system-wide implementation, both of which are addressed in later chapters in this book.

To sum up, the critiques by Bennett and others raise important points about the gaps in our knowledge and, in particular, the complexity of integrating different aspects of expertise in real work contexts. Like researchers everywhere, Bennett (2011) and Kingston and Nash (2011) suggest that more research is needed. Of course, more research is *always* needed, and this is not just a self-serving plea by researchers for continued employment. Research, particularly in the social sciences, tends to reveal that the things we study are much more complex than we assumed them to be.

However, concluding that more research is needed is not an option for school leaders. For students, schooling is basically a one-shot deal, and leaders have to make decisions right now about where to invest their efforts in supporting the teachers they lead. More evidence about what works would be nice, but right now, it simply isn't there. As we saw in the previous chapter, systematic reviews of research are simply not capable, currently, of providing reliable guides to action, and so we have to go with the evidence we have. There are significant problems with the evidence about formative assessment, but right now, there appears to be nothing that leaders can prioritise that would have a greater impact on student achievement. This is important because it means that if the only priority is raising scores on standardised tests, developing teachers' practice of formative assessment is likely to be the most effective strategy.

This is important also because many teachers say to me, "I'd love to teach for deep understanding, but I have to raise my students' test scores." What the research on formative assessment shows is that you don't have to choose. Developing the use of formative assessment is the best way to teach for deep understanding and, at the same time, the best way to improve test scores. The reason this matters is because there is a low road and a high road to test success. Taking the low road means drilling students to become better test takers. This can be effective for some students, such as those with good memories, but it is on average ineffective because such an approach de-skills students. When faced with a test item that they do not know how to answer, they tend to try to remember what they have been taught about such items, rather than thinking hard about the item that is in front of them. However, the real damage of taking the low road is that even for students who are successful, it is ineffective in the longer term. They can pass the test, but that's all they can do. Taking the high

road means teaching for deep understanding so that the students can be successful on standardised tests *and* have capabilities on which they can build in the future.

Approaches to Formative Assessment

This chapter has adopted a deliberately broad approach to formative assessment to be inclusive and head off disputes about what is and is not formative assessment, because such disputes are unlikely to be productive. However, adopting a broad definition of formative assessment does not mean that all the different approaches that such a broad definition would include are likely to be equally effective in improving student achievement.

As Shepard (2010) points out, the research evidence reviewed in recent years very strongly indicates that the approaches to formative assessment that are likely to have the greatest impact on student achievement "typically involve much more immediate interactions between students and teachers during the course of a lesson or unit of study" (p. 247). This does not mean that other approaches to formative assessment cannot be effective. For example, when there is poor alignment between curriculum and instruction on the one hand and assessment on the other, then assessments that help teachers and administrators detect and address this can be helpful (Goe & Bridgeman, 2006).

Perhaps the most widespread approach to formative assessment currently in use, however, is focused on what might be called "instructional data teams" in which teachers meet regularly to review evidence on the progress of students, and where students are found not to be making adequate progress, appropriate interventions are made.

According to Richard DuFour (2004), three ideas are central to this approach. The first is that all students should be learning. While this is common as an aspiration, making it a reality involves substantial changes to the typical operating procedures of most schools. As DuFour himself says, "Don't tell me you believe all children can learn; tell me what you do when they don't" (Blackburn, 2014). It involves re-engineering the school and its operating procedures to ensure that all students learn.

The second central idea is that teachers work collaboratively to solve problems. In many high-reliability organisations, significant advances have been made by treating failures as system failures rather than as failures of individuals (Roberts, 1990), and similar approaches appear to be highly effective in schools (Stringfield, 1995). In the same way that hospitals seek to improve the reliability of their processes, schools can do the same through collaboration:

> The powerful collaboration that characterises professional learning communities is a systematic process in which teachers work together to analyse and improve their classroom practice. Teachers work in teams, engaging in an ongoing cycle of questions that promote deep team learning. This process, in turn, leads to higher levels of student achievement. (DuFour, 2004, p. 9)

The third central idea is that the work of professional learning communities must be focused on outcomes for students, assessed through common measures. This last point is essential. When teachers are free to determine their own learning outcomes for students, then the fact that some teachers' students do less well than other students does not, in itself, force those teachers to confront the issue. They can respond by saying that the results are lower than those of other teachers because other teachers have lower standards. However, when teachers collaboratively develop common assessments, which they agree encapsulate what they want their students to learn, then differences in student success generate much more focused discussions. In particular, when the results of the common assessments are broken down to identify different aspects of success, teachers can see which of their colleagues have been particularly successful in developing those particular aspects with their students and are thus likely to be ready to learn from their colleagues.

A study of schools in a large urban school district compared student achievement in nine schools where teachers engaged in level-based instructional data teams with that in six matched schools that did not (Saunders, Goldenberg & Gallimore, 2009). In the first phase of the "Getting Results" (GR) project, which lasted two years, school principals met monthly with the regional head for two hours, and in these meetings, principals were introduced to, and supported in, the role they were expected to play in leading change in their schools. Specifically, principals were expected to establish an instructional leadership team (ILT) for the building, consisting of at least one representative from each year level, the principal, and other relevant individuals at the school (administrators, coaches, coordinators, etc.), which would meet once each month for two hours.

The task of the ILT was to support the year-level representatives in their work of leading weekly, level-based teams that would examine student work to identify academic problems and indicators of progress. Unfortunately, over the first two years of the project, it quickly became clear that the project was not meeting its intended goals.

> However, despite continuing support by the regional [head] and favourable principal responses to the GR intervention, the Phase 1 yielded limited implementation, minimal implementation or impact of any kind, and no appreciable gains in student achievement. Competing demands for their

> time and attention were typically cited as reasons for the lack of progress in implementation. Principals expressed uncertainty about the content or structure of ILT meetings and how they should lead or guide ILT representatives, who were in turn expected to lead their colleagues at [year-level] team meetings. It turned out to be very difficult for ILT representatives to function effectively as [year-level] leaders for instructional improvement, and principals were challenged to provide them with the necessary guidance. It became clear that a "train the principal" approach yielded little implementation, ineffective teacher teams, or no gains in student achievement. (Saunders et al., 2009, p. 10)

In the second phase of the project, therefore, the amount of support was increased and focused more directly on supporting the work of the various teams. Principals continued to meet monthly with project advisors, but now project advisors also met individually with principals at their schools to discuss the progress of ILTs and year-level teams. Project advisors also attended meetings of ILTs and, when requested, some year-level team meetings.

Participating teachers also received training on analysing standardised and periodic assessments, unit and instructional planning, and focusing on and addressing common student needs. A specific seven-step protocol for achieving this, developed by project staff, was given to all year-level teams:

1. Identify and clarify specific and common student needs to work on together.
2. Formulate a clear objective for each common need and analyse related student work.
3. Identify and adopt a promising instructional focus to address each common need.
4. Plan and complete necessary preparation to try the instructional focus in the classroom.
5. Try the team's instructional focus in the classroom.
6. Analyse student work to see if the objective is being met and evaluate the instruction.
7. Reassess: Continue and repeat cycle or move on to another area of need. (Saunders et al., 2009, p. 11)

The results were dramatic. While there had been no improvement in the first phase of the project, during the three years of the second phase, compared with the

matched schools who had chosen an alternative school-improvement plan, student achievement increased by approximately 0.2 standard deviations.

This study provides clear evidence that having teachers engaging in an inquiry cycle where students' needs are identified, and solutions are then proposed and developed, can be effective. However, the major strength of this model is also a weakness, and that is that it is focused on the progress of this year's students. Making sure that this year's students succeed is of course important. But if the only outcome of this process is higher achievement for this year's students, then the same problems are likely to arise in future years, because the teachers will not have improved. Now of course it could be that engaging teachers in problem solving to address the issue of students not making the progress that they need to make does result in increasing the quality of the teachers. However, if such improvements happen, they are a by-product of the process rather than the main focus. So although time has to be found to ensure that all students are learning, time also has to be found for teachers to increase their classroom skills, so that next year, fewer students fail to progress, because the quality of instruction they have received is higher. And of course, given the research on formative assessment discussed earlier in this chapter, the most obvious focus for improving teacher quality is to focus on classroom formative assessment.

Investing in increases in teacher quality is particularly powerful as a mechanism for school improvement because once a teacher gets better, even if only slightly, the benefits of improved teacher quality are experienced by every student that teacher will ever teach. And if, the following year, the teacher gets better again, all the students that teacher will teach get the double benefit of the improvements that the teacher made in the first year and the benefits of the improvements that the teacher makes in the second year – rather like the effects of compound interest.

These two processes – making sure that this year's students are making appropriate progress and making sure that teachers have time to develop their classroom skills – are complementary. One way of thinking about the relationship between the two processes is in terms of the distinction between quality control and quality assurance.

Quality control is generally defined as a set of activities undertaken by manufacturers for ensuring that whatever they release is free from defects. As a process, quality control tends to focus on checking the quality of finished products at the end of the production process and before they are released. It is therefore mainly a reactive process.

Quality assurance is generally defined as a set of activities for ensuring that the processes used in manufacturing are designed so as to eliminate, or at least minimise, defects in the final product. As a process, quality assurance tends to focus on the design of production processes and is therefore primarily a proactive process.

In much of the management literature, it is common to find quality control being regarded as inferior to quality assurance. Quality assurance is sometimes described as designing quality into the production process, while quality control is described as inspecting quality in the final stage. Advocates of quality assurance often compare the traditions of automobile manufacturing in the United States and Japan in the 1960s. In the United States, the dominant approach was quality control; cars were manufactured and then inspected. Cars that failed the inspection were sent back to have the defects corrected. Beginning in the late 1950s, building on the work of Americans such as W. Edwards Deming, Japanese manufacturers – most notably the Toyota Motor Corporation – began to focus on quality assurance. The production process would be designed so that quality was engineered into each step of the production process. Every worker was given the right (and, indeed, the responsibility) to halt the production line if any issue that might affect the quality of the final product was seen.

Today, of course, all automobile manufacturers prioritise quality assurance as the central approach to ensuring quality, but it is important to realise that quality assurance cannot be the exclusive approach to quality, especially when what is being produced is highly complex. For example, many manufacturers of semiconductor chips have found that it is actually very difficult to design manufacturing processes so that every chip that is produced is guaranteed to work properly. Obviously they take care to design the manufacturing process so as to maximise the chances of success, but the optimum strategy consists of both quality assurance *and* quality control – making sure that the production process is well designed but also checking on the quality of the finished product.

Instructional data teams can be thought of as a process of quality control. The idea is that student achievement is assessed after a sequence of instruction has been completed, with a view to taking action if certain students are found not to have learned what they have been taught. In keeping with the definition of quality control mentioned previously, it is primarily a reactive process. In contrast, getting teachers to make greater use of classroom formative assessment can be thought of as a process of quality assurance. It is about designing instruction so that any failures of students to learn are identified and can be addressed while the instruction is still taking place. Both processes are necessary because, as Kirschner, Sweller and Clark (2006) point out, "The aim of all instruction is to alter long-term memory. If nothing has changed in long-term memory, nothing has been learned" (p. 77). However, while we cannot be sure that students have learned something because they know it at the end of the lesson, if they do not know something by the end of the lesson, it is unlikely that they will know it weeks later. A summary of the main differences between these two approaches to formative assessment is shown in Table 4.2.

Table 4.2. Comparison of Two Main Approaches to Formative Assessment

Improving student achievement with instructional data teams	Building teacher quality through classroom formative assessment
Quality control	Quality assurance
Common assessments	Formative-assessment strategies and techniques
Improvement through better teamwork and systems	Improvement through increased teacher capability
Focus on individual outcomes for students	Focus on teachers' individual accountability for change
Regular meetings focused on data	Regular meetings focused on teacher improvement
Sixteen points on PISA (in two to three years)	Thirty points on PISA (in two to three years)

Given that leaders need to ensure that both of these processes are in place, the obvious question is how to divide time between the two approaches. The good news, as we shall see in chapter 6, is that there is not much point in having meetings focused on building teacher capability more frequently than once a month. Teachers' lives and working routines are so hectic and fragile that it takes most teachers a month to try out a new teaching technique. This suggests that if the school schedule allows time for teachers to work collaboratively once a week, then three meetings per month should be focused on instructional data, and one meeting a month should be focused on building teacher capability.

Furthermore, to make clear the difference between the two kinds of meetings, it is helpful if instructional data teams meet in *level-based* teams, and meetings focused on building teacher capability involve *cross-level* teams. Having teachers meet in cross-level teams minimises the likelihood that teachers spend time talking about curriculum or students, and as a result, they are more likely to spend time on the one thing that all teachers have in common – pedagogy. Detailed guidance about the structure and organisation of meetings designed to support teachers in their use of classroom formative assessment is provided in chapter 6.

Relationships With Other Policy Priorities

So far in this chapter, I have argued that leaders who are serious about improving the outcomes for students in their schools have to develop the use of formative assessment, both retrospectively, as a way of ensuring that students do not fall behind, and also prospectively, as a way of increasing the pedagogical skills of teachers in the school. However, schools rarely have the luxury of focusing on a single priority. Particularly in recent years, states that have secured federal funds through the Race to the Top program have been required to implement rigorous teacher evaluation systems, which raises issues about the extent to which developing formative assessment is consistent with the common teacher evaluation systems being adopted.

Perhaps the most important point to remember in looking at teacher evaluation systems is that, by their very nature, they are designed to be comprehensive. They have to include all aspects of teacher practice that are likely to be relevant, no matter how important they are in contributing to student progress. We saw in chapter 2 (Sartain et al., 2011) that in Chicago Public Schools, there were significant correlations between the progress made by students and a teacher's rating on two of the four domains of Danielson's Framework for Teaching (Danielson, 1996): domain 2 (classroom environment) and domain 3 (instruction). However, no such correlation was found for domains 1 (planning and preparation) and 4 (professional responsibilities). As I said earlier, this does not mean that planning and preparation and professional responsibilities are unimportant aspects of teachers' performance. It could be that they are so important that all teachers take them equally seriously, so that teachers do not vary much in terms of these two domains, so the correlation is reduced. It could be that variations in terms of these two domains are important, but the definitions of the different levels of performance do not capture these differences. Finally, of course, it could be that the variations in student progress in these two domains are smaller than those produced by variations in classroom environment and instruction.

Whatever the reasons, a teacher evaluation system that did not cover planning and preparation and professional responsibilities would lack credibility, and so these dimensions of teacher performance have to be included. However, the empirical evidence is that these matter less than classroom environment and instruction. Put bluntly, it seems likely that if we want to improve a teacher's performance, attention to classroom environment and instruction will benefit students more than attention to planning and preparation and professional responsibilities.

The same argument can be made *within* each domain. While each domain is specified in terms of a number of subdomains, it is likely that certain subdomains will be more important than others (in the sense that improvement of that aspect of practice produces bigger improvements in student achievement than improvements in the

other aspects of practice). Now of course, improving practice in some subdomains may be harder than others, so the fact that a particular aspect of practice has the highest correlation with student achievement should not necessarily lead us to focus on that aspect. We need instead to focus on aspects of practice that can be changed relatively easily and also, when changed, have a major impact on student achievement.

What this means is that it is highly unwise to use a teacher *evaluation* framework as a teacher *improvement* framework. Teacher evaluation frameworks have to be comprehensive and, specifically, they need to cover all aspects of teachers' work, so they include some aspects of practice that have a huge impact on student learning, and they also include aspects of practice that have small, or negligible, impacts on student learning. And yet, in almost all evaluation frameworks, all aspects of practice carry equal weight. What this means is that when teachers are under pressure to improve their ratings, the incentives to improve aspects of practice that have no impact on student achievement are the same as improving aspects of practice that really benefit their students. When evaluation frameworks are used as improvement frameworks, we create incentives for teachers to focus on the easiest things to improve, rather than what will benefit their students most. Put bluntly, *using evaluation frameworks as improvement frameworks makes it more likely that we improve teachers in ways that do not benefit their students*.

Fortunately, the research base on formative assessment provides a way of focusing teacher improvement so as to maximise the impact on student achievement. Figure 4.2 shows an illustration of the connections between the five strategies of classroom formative assessment presented in Figure 4.1 and the five subdomains of domain 3 in Danielson's Framework for Teaching.

To unpack the rather dense interconnections shown in Figure 4.2, the aspect of "communicating with students" within the Framework for Teaching that are given particular emphasis in formative assessment are the ideas of sharing learning intentions and success criteria with students and providing feedback that moves learning forward. What is called "using questioning and discussion techniques" in the Framework for Teaching maps in a fairly straightforward way to the strategy of "eliciting evidence" in classroom formative assessment. Aspects of what is called "engaging students in learning" in the Framework for Teaching that are a particular focus of classroom formative assessment are sharing learning intentions, activating students as learning resources for one another, and activating students as owners of their own learning. The last two subdomains of the Framework for Teaching, using assessment in instruction and demonstrating flexibility and responsiveness, of course map to all five of the strategies of classroom formative assessment. Indeed, it could be argued that the entire purpose of classroom formative assessment is to demonstrate flexibility

and responsiveness, and it is only possible to demonstrate flexibility and responsiveness by using assessment as an integral part of instruction.

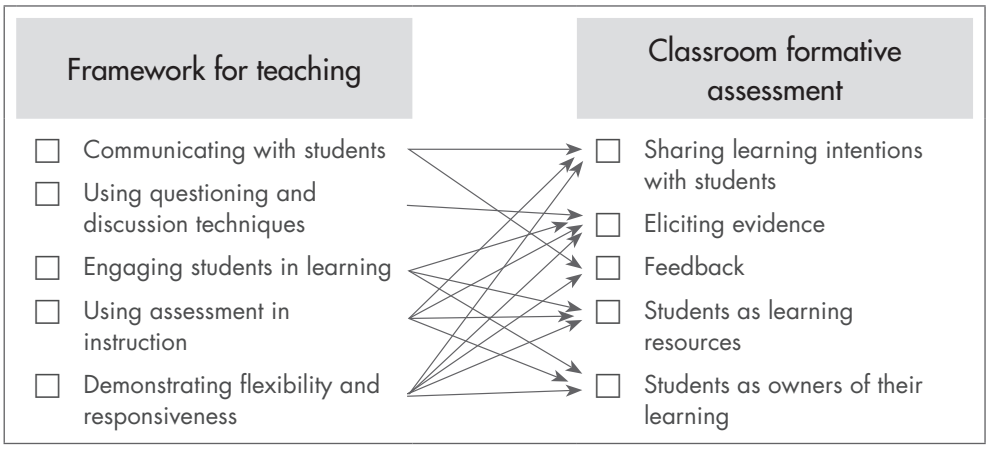

Figure 4.2. Connections between the Framework for Teaching and classroom formative assessment

In this way, classroom formative assessment can be thought of as placing a magnifying glass on the aspects of the Framework for Teaching that are likely to have the greatest impact on student achievement. To see why this really is a magnifying glass, refer back to where we saw earlier that improving a teacher from "below basic" to "distinguished" would equate to a 30 per cent increase in the rate of learning, and such a process is likely to take several years at the very least (if it can be done). Getting teachers to develop their use of classroom formative assessment appears to result in a 50 per cent increase in the rate of student learning within a year or two.

Other frameworks divide up the work of teachers differently, but the important point is that any comprehensive teacher evaluation framework will include all aspects of classroom formative assessment somewhere in the framework. This means that leaders can be confident that encouraging teachers to focus on classroom formative assessment will allow teachers to make progress with respect to the evaluation framework, which is obviously in the teachers' best interest. The leaders can also be confident that the aspects that are being developed are those that are most likely to benefit students, thus reducing the likelihood of improvements in teacher evaluations that are not associated with improvements in student achievement.

At this point, it is also worth noting that classroom formative assessment is an inherent element of two other important initiatives of recent years, differentiated instruction and response to intervention (and instruction). Each of these is discussed in turn.

Differentiated Instruction

Although the idea of differentiated instruction has received attention only in recent years, it is important to realise that there is nothing new in the idea. In several countries (e.g. Germany, Austria), the school performance of students is evaluated, and on the basis of that evaluation, students are directed to different kinds of educational institutions where they are expected to learn different things. In other words, different students are assigned different goals in education. In other jurisdictions, students might work toward the same goals but do so with different curriculum structures, different course content, different tasks or different teaching approaches. Perhaps most common is the idea of differentiation in the pace of learning, with certain students being expected to cover course content more quickly than others, after which they might go on to work normally intended for older students (acceleration) or to undertake more in-depth study of the work already completed (enrichment). Finally, of course, particularly as a result of mandated testing of students, there are also differences in the way students are assessed, with some students receiving accommodations and adaptations of the assessment administered.

There does not appear to be any widely accepted definition of the term *differentiated instruction*. Carol Ann Tomlinson (2004) – who is the writer most closely associated with the idea of differentiated instruction – has suggested that

> While the concept of "differentiated instruction" can be defined in many ways, as good a definition as any is ensuring that what a student learns, how he/she learns it, and how the student demonstrates what he/she has learned is a match for that student's readiness level, interests and preferred mode of learning. (p. 188)

Differentiated instruction has been widely adopted as a key element of instructional policy in many school systems, even though convincing evidence that it improves learning is rather difficult to locate (Schmoker, 2010). Partly, this is due to the lack of an agreed-upon definition, but even with such a definition, some aspects of differentiated instruction are simply unresearchable. For example, if students exercise choice in what they study, we cannot compare student achievements. In Robert Wood's memorable phrase, it makes no sense to say, "Your chemistry equals my French" (Wood, 1987). Similarly, if different students are assessed on different bases, meaningful comparisons are difficult, if not impossible.

When students are learning the same material and are assessed in the same way, then, as David Ausubel has pointed out in the quotation earlier in this chapter, it makes sense to start from where the learner is. Beginning instruction by assuming students know something they do not know is hardly likely to be a recipe for effective instruction. And even if students are at different starting points, what the teacher

should do with that is determined by cultural as much as by scientific considerations. Typically, it seems to be generally assumed that those who have learned something should move ahead of their peers, while in other cultures, students who have learned something are expected to help those in the class who haven't.

In a review of the research evidence about differentiated learning, Huebner (2010) states that "experts and practitioners acknowledge that the research on differentiated instruction as a specific practice is limited" (p. 79) but then goes on to suggest that there is research to support a number of practices that provide the foundation of differentiated instruction:

> using effective classroom management procedures; promoting student engagement and motivation; assessing student readiness; responding to learning styles; grouping students for instruction; and teaching to the student's zone of proximal development (the distance between what a learner can demonstrate without assistance and what the learner can do with assistance). (p. 79)

The problem with this claim is that the first three of these would probably be regarded as aspects of *all* effective instruction, and the last three are, at best, questionable. As we saw in chapter 3, there is no evidence that adapting instruction to mesh with students' preferred ways of learning has any impact on student achievement. Indeed, catering to students' preferred mode of learning may actually be suboptimal for student learning because of what Elizabeth and Robert Bjork (2009) have called "desirable difficulties" in learning. Within-class grouping of students for instruction can produce small improvements in student learning (Lou et al., 1996), but many studies have found that teachers end up spending up to half their time keeping groups engaged (e.g. Good, Grouws, Mason, Slavings & Cramer, 1990). Finally, the characterisation of Vygotsky's zone of proximal development does not reflect the careful distinction that Vygotsky drew between learning and development. Like Piaget, Vygotsky believed that children went through a series of stages of psychological development. For Vygotsky (1978), the purpose of teaching was to trigger a transition from one stage of development to the next, which is why he wrote, "Thus, the notion of a zone of proximal development enables us to propound a new formula, namely that the only 'good learning' is that which is in advance of development" (p. 34).

Therefore, while differentiated instruction has neither an agreed-upon definition nor a coherent research foundation, leaders are often required to ensure that the instruction in their schools is differentiated. A way forward is suggested by the work of Hall, Strangman and Meyer (2011) who define differentiated instruction in a similar way to Tomlinson:

> To differentiate instruction is to recognise students' varying background knowledge, readiness, language, preferences in learning and interests; and to react responsively. Differentiated instruction is a process to teaching and learning for students of differing abilities in the same class. (p. 3)

While this definition does not, in itself, add much to the earlier definition from Tomlinson, what is helpful in Hall et al.'s report (2011) is that they present a list of thirteen characteristics of differentiated instruction, and these are shown in Table 4.3. Also included in Table 4.3, in the final column, is an indication of whether each characteristic is also an aspect of classroom formative assessment. By encouraging teachers to develop their use of those characteristics of differentiated instruction that are also aspects of classroom formative assessment, leaders can maximise the chances that attention to differentiated instruction will also help learners.

Table 4.3. Aspects of Differentiated Instruction That Are Related to Classroom Formative Assessment

	Aspects of differentiated instruction	Formative assessment?
Content	Several elements and materials are used	
	Align tasks and objectives to learning goals	√
	Instruction is concept focused and principle driven	
Process	Flexible grouping is consistently used	
	Classroom management benefits students and teachers	
Products	Initial and ongoing assessment of student readiness and growth	√
	Students are active and responsible explorers	√
	Vary expectations and requirements for student responses	√
Miscellaneous	Clarify key concepts and generalisations	
	Use assessment as a teaching tool	√
	Emphasise critical and creative thinking as a goal in lesson design	
	Engaging all learners is essential	√
	Balance between teacher-assigned and student-selected tasks	

Response to Intervention (and Instruction)

The field of education is famous (or infamous) for its use of impenetrable jargon, but even so, the phrase *response to intervention* must be one of the most impenetrable and least intuitive descriptions to have appeared in recent years.

The origin of the phrase lies with the American Individuals with Disabilities Education Improvement Act of 2004, which made a number of changes to methods for determining whether children had learning disabilities that had been implicit in the earlier Individuals with Disabilities Education Act of 1990.

Traditionally, specific learning disabilities had been diagnosed by reference to differences in aspects of cognitive performance. For example, as Siegel (2006) notes:

> Until recently, the typical definition of dyslexia involved a discrepancy between an IQ score and a reading score. If the IQ score was found to be significantly higher than the reading score, then this discrepancy was used as an index of dyslexia. (p. 582)

However, this approach has been subject to a number of criticisms. It appears that the use of IQ testing is not particularly helpful in identifying specific learning disabilities, such as dyslexia (e.g. Siegel, 1989). Perhaps more important, given the association of IQ testing with the eugenics movement in the first half of the 20th century (Selden, 1999), it is not surprising that those tasked with updating the 1990 Individuals with Disabilities in Education Act sought to avoid the use of IQ testing in the diagnosis of learning disabilities. Whatever the reason, section 614 of the 2004 act states:

> When determining whether a child has a specific learning disability … a local educational agency shall not be required to take into consideration whether a child has a severe discrepancy between achievement and intellectual ability in oral expression, listening comprehension, written expression, basic reading skill, reading comprehension, mathematical calculation, or mathematical reasoning. (Title I/B/ § 614/b/6)

Instead, local education agencies are encouraged to adopt more direct ways of diagnosing learning disabilities, and the most obvious of all is that a student may have a learning disability if she or he fails to learn. The difficulty with such a definition is of course that a failure to learn may simply indicate ineffective instruction. That is why section 614 of the 2004 act also states:

> In determining whether a child has a specific learning disability, a local educational agency may use a process that determines if the child responds

> to scientific, research-based intervention as a part of the evaluation procedures. (Title I/B/ § 614/b/6)

In other words, if a student does not *respond* to a research-based *intervention*, then the student may, indeed, have a learning disability.

Although the origins of the phrase *response to intervention* therefore lie in the desire to avoid requiring local education agencies to use IQ testing in the identification of learning disabilities, as it has become more widely used and applied, it is now being used much more generally as a protocol for preventing academic failure by monitoring progress and, when evidence reveals that students are not making progress, taking action – in essence, a way of implementing the quality control approach to formative assessment described previously. As defined by the American National Center on Response to Intervention (2010):

> Response to intervention integrates assessment and intervention within a multi-level prevention system to maximise student achievement and reduce behaviour problems. With RTI, schools identify students at risk for poor learning outcomes, monitor student progress, provide evidence-based interventions and adjust the intensity and nature of those interventions depending on a student's responsiveness, and identify students with learning disabilities.

The relationship with classroom formative assessment should now be clear. To maximise student achievement, the regular instruction provided to all students (often called "tier 1 instruction" within RTI approaches) must be as good as it can be, and this requires the use of regular classroom formative assessment. Formative assessment also involves assessing the progress made by learners, and where this is deemed inadequate, more intensive interventions (tier 2 and tier 3) are used. Moreover, because formative assessment, at least when it is done well, indicates not just that students are not learning but why and what can be done about it, then formative assessment actually increases the effectiveness of tier 2 interventions because they can be more effectively targeted to the specific difficulties the student is experiencing.

The purpose of these rather extended discussions into the Framework for Teaching, differentiated instruction, and response to intervention has been to show that classroom formative assessment is not just compatible with these initiatives. Rather, formative assessment becomes the vehicle for *delivering* them, and not just an addition to the load that teachers and leaders have to bear. By focusing on the aspects of the Framework for Teaching that have the greatest impacts on student achievement, classroom formative assessment aligns the attempts of teachers to improve their ratings with improved outcomes for their students, which might not

otherwise happen. By highlighting the aspects of differentiated instruction that are supported by evidence, leaders can ensure that compliance with mandates that differentiated instruction should be in place will actually benefit learners. And finally, this chapter has shown that the whole idea of response to intervention is just one aspect of classroom formative assessment. The pieces really do fit together.