2

AN INTEGRATIVE SUMMARY OF THE RESEARCH LITERATURE AND IMPLICATIONS FOR A NEW THEORY OF FORMATIVE ASSESSMENT

DYLAN WILIAM

If what students learned as a result of a particular sequence of instruction was predictable, there would be no need for assessment. Educators could just compile an inventory of what they had taught and use this inventory as a catalogue of what students had learned. This was, in effect, the underlying assumption of the educational model in the medieval English universities of Oxford and Cambridge, where a bachelor's degree was conferred after the completion of a certain period of residence. Of course, as research studies (e.g., Denvir & Brown, 1986a, 1986b)—and the experience of educators—attest, what students learn from a particular sequence of instruction can be very different from what the teacher intended to teach them. That is why assessment is a central and perhaps even a defining feature of effective instruction: Assessment is the only way that we can know whether what has been taught has been learned. In a very real sense, therefore, assessment is the bridge between learning and teaching.

Assessment is what makes the routine coming together of teachers and students for the purpose of creating learning different from, for example, that of a teacher speaking into a video camera that is then transmitted to students in another room: Together, teachers and students can ensure that information about student achievement, gained through assessment, can be used to adjust the instruction in order to better meet student learning needs. This is the essence of formative assessment: the idea that evidence of student achievement is elicited, is interpreted, and leads to action that results in better learning than would have been the case in the absence of such evidence (Wiliam & Black, 1996).

The origins of the term *formative assessment* have been detailed elsewhere (see Cizek, this volume; Guskey, this volume; Wiliam, 2007a). The aim of this chapter is to build on the basic idea of formative assessment to try to provide a clear theoretical basis for the ways in which assessment can support learning, to show how the various formulations of the notion of formative assessment that have been proposed over the last 40 years can be encompassed within a broader overarching framework, and to indicate briefly how that framework connects to research in related areas.

REVIEWS OF RESEARCH ON FEEDBACK AND FORMATIVE ASSESSMENT

One of the powerful metaphors that underlie the theory of action of formative assessment is the idea of *feedback*, developed originally in the field of systems engineering (see Wiener, 1948). As Ramaprasad (1983) noted, the defining feature of feedback is that the information generated within the system must have some effect on the system. Information that does not have the capability to change the performance of the system is not feedback. Ramaprasad said: "Feedback is information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way" (p. 4). Commenting on this, Sadler (1989) noted:

> An important feature of Ramaprasad's definition is that information about the gap between actual and reference levels is considered as feedback only when it is used to alter the gap. If the information is simply recorded, passed to a third party who lacks either the knowledge or the power to change the outcome, or is too deeply coded (for example, as a summary grade given by the teacher) to lead to appropriate action, the control loop cannot be closed, and "dangling data" substituted for effective feedback. (p. 121)

In this view, feedback cannot be separated from its instructional consequences. It is therefore not surprising that over the last quarter century, a number of substantial reviews have appeared concerning the impact of assessment practices on students and their learning in the context of the classroom (Allal & Lopez, 2005; Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Black & Wiliam, 1998a, 1998b; Brookhart, 2004, 2007; Crooks, 1988; Dempster, 1991, 1992; Elshout-Mohr, 1994; Fuchs & Fuchs, 1986; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Köller, 2005; Natriello, 1987; Nyquist, 2003; Shute, 2008; Wiliam, 2007a).

The reviews resist any easy synthesis due to differences in their starting assumptions, their theoretical bases, and their remits, and besides, a detailed summary of each of these reviews is beyond the scope of this chapter (Brookhart, 2004). Nevertheless, some significant themes emerge.

The first theme is that the outcomes of assessment are used in a multiplicity of ways, with different uses that are often in conflict (Black & Wiliam, 1998a; Crooks, 1988; Natriello, 1987). In particular, the use of assessments for summative purposes (such as determining a grade on a course) appears to reduce the extent to which they can serve to support learning.

The second common theme is that different kinds of feedback may be differentially effective for different kinds of learning. For example, the kinds of feedback that are most effective in developing lower-level skills and content knowledge may not be the most effective for higher-order skills (Dempster, 1991, 1992; Elshout-Mohr, 1994), and in particular, that immediate feedback appears to be more effective for procedural learning, while delayed feedback may be more effective for higher-order outcomes (Shute, 2008).

The third, and perhaps most important, theme is that the most effective feedback focuses attention prospectively rather than retrospectively. The important question is not, "What did I get right and what did I get wrong?" but, "What next?" (Bangert-Drowns

et al., 1991; Fuchs & Fuchs, 1986; Hattie & Timperley, 2007; Nyquist, 2003). Short-term studies can be particularly misleading in this respect, because while certain kinds of feedback interventions—defined by Kluger & DeNisi (1996, p. 255) as "actions taken by (an) external agent(s) to provide information regarding some aspect(s) of one's task performance"—can increase performance, they may do so by changing the kind of motivation. For example, a feedback intervention may show positive effects by increasing task motivation, but then future learning would require continuous feedback. Even where the emphasis is on task-learning processes, feedback interventions may encourage shallow learning, thus making higher-order goals more difficult to achieve (Kluger & DeNisi, 1996; Shute, 2008).

EFFECT SIZES IN REVIEWS OF RESEARCH ON FORMATIVE ASSESSMENT AND THEIR LIMITATIONS

The reviews of research cited above produce a range of estimates of the size of the effect that the use of formative feedback might be expected to have on learning. Bangert-Drowns et al. (1991), found an average effect of around one-fourth of a standard deviation for feedback in testlike events, while Kluger and DeNisi (1996) and Nyquist (2003) found that feedback produced larger effect sizes—around 0.4 standard deviations—although both noted that the variability across different studies was extremely high. Black and Wiliam (1998a) and Shute (2008) suggested that typical effect sizes were in the range 0.4 to 0.7 and 0.4 to 0.8 respectively while a review of 74 meta-analyses of the effects of feedback by Hattie and Timperley (2007) found an average effect size of 0.95 standard deviations across 4,157 studies.

The use of standardized effect sizes to compare and synthesize studies is understandable, because few of the studies included in the various reviews published sufficient details to allow more sophisticated forms of synthesis to be undertaken, but relying on standardized effect sizes in educational studies creates substantial difficulties of interpretation, for two reasons.

First, as Black and Wiliam (1998a) noted, effect size is influenced by the range of achievement in the population. An increase of 5 points on a test where the population standard deviation is 10 points would result in an effect size of 0.5 standard deviations. However, the same intervention when administered only to the upper half of the same population, provided that it was equally effective for all students, would result in an effect size of over 0.8 standard deviations, due to the reduced variance of the subsample. An often-observed finding in the literature—that formative assessment interventions are more successful for students with special educational needs (for example in Fuchs & Fuchs, 1986)—is difficult to interpret without some attempt to control for the restriction of range, and may simply be a statistical artifact.

The second and more important limitation of the meta-analytic reviews is that they fail to take into account the fact that different outcome measures are not equally sensitive to instruction (Popham, 2007). Much of the methodology of meta-analysis used in education and psychology has been borrowed uncritically from the medical and health sciences, where the different studies being combined in meta-analyses either use the same outcome measures (e.g., 1-year survival rates) or outcome measures that are rea-

sonably consistent across different settings (e.g., time to discharge from hospital care). In education, to aggregate outcomes from different studies it is necessary to assume that the outcome measures are equally sensitive to instruction.

It has long been known that teacher-constructed measures have tended to show greater effect sizes for experimental interventions than obtained with standardized tests, and this has sometimes been regarded as evidence of the invalidity of teacher-constructed measures. However, as has become clear in recent years, assessments vary greatly in their sensitivity to instruction—the extent to which they measure the things that educational processes change (Wiliam, 2007b). In particular, the way that standardized tests are constructed reduces their sensitivity to instruction. The reliability of a test can be increased by replacing items that do not discriminate between candidates with items that do, so items that all students answer correctly, or that all students answer incorrectly, are generally omitted. However, such systematic deletion of items can alter the construct being measured by the test, because items related to aspects of learning that are effectively taught by teachers are less likely to be included than items that are taught ineffectively.

For example, an item that is answered incorrectly by all students in the seventh grade and answered correctly by all students in the eighth grade is almost certainly assessing something that is changed by instruction, but is unlikely to be retained in a test for seventh graders (because it is too hard), nor in one for eighth graders (because it is too easy). This is an extreme example, but it does highlight how the sensitivity of a test to the effects of instruction can be significantly affected by the normal processes of test development (Wiliam, 2008).

The effects of sensitivity to instruction are far from negligible. Bloom (1984) famously observed that one-to-one tutorial instruction was more effective than average group-based instruction by two standard deviations. Such a claim is credible in the context of many assessments, but for standardized tests such as those used in the National Assessment of Educational Progress (NAEP), one year's progress for an average student is equivalent to one-fourth of a standard deviation (NAEP, 2006), so for Bloom's claim to be true, one year's individual tuition would produce the same effect as 9 years of average group-based instruction, which seems unlikely. The important point here is that the outcome measures used in different studies are likely to differ significantly in their sensitivity to instruction, and the most significant element in determining an assessment's sensitivity to instruction appears to be its distance from the curriculum it is intended to assess.

Ruiz-Primo, Shavelson, Hamilton, and Klein (2002) proposed a five-fold classification for the distance of an assessment from the enactment of curriculum, with examples of each:

1. *Immediate*, such as science journals, notebooks, and classroom tests;
2. *Close*, or formal embedded assessments (for example, if an immediate assessment asked about number of pendulum swings in 15 seconds, a close assessment would ask about the time taken for 10 swings);
3. *Proximal*, including a different assessment of the same concept, requiring some transfer (for example, if an immediate assessment asked students to construct

boats out of paper cups, the proximal assessment would ask for an explanation of what makes bottles float or sink);
4. *Distal*, for example a large-scale assessment from a state assessment framework, in which the assessment task was sampled from a different domain, such as physical science, and where the problem, procedures, materials and measurement methods differed from those used in the original activities; and
5. *Remote*, such as standardized national achievement tests.

As might be expected, Ruiz-Primo et al. (2002) found that the closer the assessment was to the enactment of the curriculum, the greater was the sensitivity of the assessment to the effects of instruction, and that the impact was considerable. For example, one of their interventions showed an average effect size of 0.26 when measured with a proximal assessment, but an effect size of 1.26 when measured with a close assessment.

In none of the meta-analyses discussed above was there any attempt to control for the effects of differences in the sensitivity to instruction of the different outcome measures. By itself, it does not invalidate the claims that formative assessment is likely to be effective in improving student outcomes. Indeed, in all likelihood, attempts to improve the quality of teachers' formative assessment practices are likely to be considerably more cost-effective than many, if not most, other interventions (Wiliam & Thomson, 2007). However, failure to control for the impact of this factor means that considerable care should be taken in quoting particular effect sizes as being likely to be achieved in practice, and other measures of the impact, such as increases in the rate of learning, may be more appropriate (Wiliam, 2007c). More importantly, attention may need to be shifted away from the size of the effects and toward the role that effective feedback can play in the design of effective learning environments (Wiliam, 2007a). In concluding their review of over 3,000 studies of the effects of feedback interventions in schools, colleges and workplaces, Kluger and DeNisi observed that:

> considerations of utility and alternative interventions suggest that even an FI [feedback intervention] with demonstrated positive effects should not be administered wherever possible. Rather additional development of FIT [feedback intervention theory] is needed to establish the circumstance under which positive FI effects on performance are also lasting and efficient and when these effects are transient and have questionable utility. This research must focus on the processes induced by FIs and not on the general question of whether FIs improve performance—look how little progress 90 years of attempts to answer the latter question have yielded. (1996, p. 278)

The remainder of this chapter reviews a number of recent definitions of formative assessment and proposes a definition of formative assessment in terms of the function that assessment evidence fulfills; specifically, the extent to which assessment supports and improves instructional decisions. The consequences of this definition are then examined, focusing in particular on how formative assessment may be operationalized, and the chapter concludes by sketching out briefly some links to other related areas of research and some priorities for future research.

DEFINITIONS OF FORMATIVE ASSESSMENT

A variety of definitions of the term *formative assessment* have been proposed over the years. In their review, Black and Wiliam (1998a) defined formative assessment "as encompassing all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged" (p. 7). In a subsequent publication, addressed to policymakers and practitioners, Black and Wiliam adopted the following definition:

> 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. (1998b, p. 140)

Cowie and Bell (1999) adopted a slightly more restrictive definition by limiting the term to assessment conducted and acted upon while learning was taking place. They defined formative assessment as "the process used by teachers and students to recognize and respond to student learning in order to enhance that learning, during the learning" (p. 32). The requirement that the assessment be conducted during learning was also embraced by Shepard, Hammerness, Darling-Hammond, and Rust (2005) in their definition of formative assessment as "assessment carried out during the instructional process for the purpose of improving teaching or learning" (p. 275). In their review of formative assessment practices across eight national and provincial systems, the Organization for Economic Cooperation and Development (OECD) also emphasized the principle that the assessment should take place during instruction: "Formative assessment refers to frequent, interactive assessments of students' progress and understanding to identify learning needs and adjust teaching appropriately" (Looney, 2005, p. 21). In a similar vein, Kahl (2005) wrote: "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" (p. 11).

Broadfoot et al. (1999) argued that improving learning through assessment depended on five key factors: (1) the provision of effective feedback to pupils; (2) the active involvement of pupils in their own learning; (3) adjusting teaching to take account of the results of assessment; (4) a recognition of the profound influence assessment has on the motivation and self-esteem of pupils, both of which are crucial influences on learning; and (5) the need for pupils to be able to assess themselves and understand how to improve.

Broadfoot et al. (1999) suggested that the term *formative assessment* was unhelpful to describe such uses of assessment because "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" (p. 7). Instead they suggested instead the use of the term *assessment for learning*.

The first use of the term assessment for learning appears to be in a paper given at the annual conference of the Association for Supervision and Curriculum Development

(James, 1992); the same year a book entitled Testing for Learning was published (Mitchell, 1992). Assessment for Learning was used as the title of a book three years later (Sutton, 1995), but the first use of the term assessment for learning as a counterpoint to assessment of learning appears to be by Gipps and Stobart (1997). The use of the term was popularized in the United Kingdom by Broadfoot et al. (1999) and in the United States by Stiggins (2002). The definition given by the Assessment Reform Group (Broadfoot et al., 2002) is: "Assessment for learning is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there" (pp. 2–3).

Whereas many authors have used the terms *formative assessment* and *assessment for learning* interchangeably, or as different labels for the same idea, Black, Harrison, Lee, Marshall, and Wiliam (2004) distinguished between the terms 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. (p. 10)

Perhaps the most important point here is the distinction between formative and summative in terms of the function the assessment serves, rather than the assessment itself. Wiliam and Black (1996) argued that attempting to use the words *formative* and *summative* to describe assessments leads to contradiction, since the same assessment instrument, and even the same assessment outcomes, could be used both formatively and summatively. While locating the distinction in terms of the purpose of the assessment overcomes some difficulties, it still leaves open the possibility that assessment evidence might be collected with the intention of supporting learning, but might never actually do so.

A NEW THEORY OF FORMATIVE ASSESSMENT: PRECISION IN DEFINITION

In order to provide a comprehensive definition of formative assessment, Black and Wiliam (2009) proposed that assessment 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. (p. 6)

In explicating this definition, Black and Wiliam (2009) elaborated on five key points. First, *anyone can be the agent in formative assessment*. Although in many cases the deci-

sions will be made by the teacher, the definition also includes those situations in which the decisions are made by the learners themselves, or their peers.

Second, *the focus of the definition is on decisions*. Black and Wiliam (2009) noted that the focus of the definition could be on the intentions of those involved in instruction in collecting the evidence, but then data collection activities that did not impact learning in any way would be potentially formative, which would be contrary to common sense (and indeed to the literal meaning of the term formative). Such a definition would, in that sense, be too open. On the other hand, the definition of Black and Wiliam (1998b) focused on the outcome. It required that the assessment did in fact lead to better learning, which would appear to be a rather stringent criterion, because there could be many situations in which actions that might be expected to increase learning might not do so, given to the unpredictable nature of learning (and students). The focus on decisions is also consistent with Alexander's definition of pedagogy as:

> the act of teaching together with its attendant discourse of educational theories, values, evidence and justifications. It is what one needs to know, and the skills one needs to command, in order to make and justify the many different kinds of decision of which teaching is constituted. (2008, p. 47)

Third, *the definition focuses on next steps in instruction*. The term *instruction* is used to describe any planful activity intended to create learning, which is here defined as an increase, brought about by experience, in the capacities of an organism to act, or react in response to stimuli, in valued ways. The term *instruction* thus subsumes the roles of both the teacher and the learner. This use of the term will be unfamiliar to some readers since the term *instruction* is used in some contexts to denote a transmissionist approach to teaching, but such a connotation is quite definitely not intended here. In this context it is worth noting that there are languages where the same word is used for both teaching and learning (Welsh: *dysgu*; Maori: *ako*). It is this inclusive sense of the word *instruction*, which denotes both teaching and learning that is intended here.

Fourth, *the definition is probabilistic*. Locating the burden of definition of the term *formative* in the resulting action creates the difficulty that proof of effect is impossible to establish, requiring the verification of a counterfactual claim: that what occurred was different (and better than) what would have happened in the absence of the assessment (but did not do so). Requiring that the decisions are likely to be better reflects the fact that even the best designed interventions will not always result in better learning for all students.

Finally, *the assessment need not change the planned instruction*. The definition requires that decisions are either better or better founded, than decisions made without the evidence elicited as part of the assessment process. The second possibility is included to include those cases where the assessment indicates to the teacher that the best course of action is in fact that which the teacher had intended prior to the elicitation of evidence. In this case, formative assessment would not change the course of action, but it would mean that it was better grounded in evidence. (On this point, thanks are due to Jim Popham, who, through relentless probing, forced a clarification of this aspect of the definition.)

From this definition, Black and Wiliam proposed that formative assessment is, in essence, concerned with "the creation of, and capitalization upon, 'moments of contingency' in instruction for the purpose of the regulation of learning processes" (2009, p. 6). A theory of formative assessment is therefore much narrower than an overall theory of teaching and learning, although it links in significant ways to other aspects of teaching and learning, since how teachers, learners, and their peers create and capitalize on these moments of contingency entails considerations of instructional design, curriculum, pedagogy, psychology, and epistemology.

Moments of contingency can be synchronous or asynchronous. Examples of synchronous moments include teachers' real-time adjustments during one-on-one teaching or whole class discussion. Asynchronous examples include teachers' feedback, the use of evidence derived from homework, or students' summaries made at the end of a lesson, each used to plan a subsequent lesson. Furthermore, these asynchronous moments might be used to modify the instruction of those from whom the evidence was collected, or the teacher may collect evidence about difficulties experienced by one group, and use this to modify instruction for another group of students at some point in the future.

Teachers' responses to information about student learning can be one-to-one or group-based. Responses to a student's written work are usually one-on-one, but in classroom discussions the feedback will be in relation to the needs of the subject-classroom as a whole, and may be an immediate intervention in the flow of classroom discussion, or a decision about how to begin the next lesson.

A NEW THEORY OF FORMATIVE ASSESSMENT: DEFINITIONAL CONSEQUENCES

In this section, two particular consequences of the definition of formative assessment just described are explored: the kinds of decisions that formative assessments can support, and the immediacy of the instructional adjustments that are informed by the assessments.

What Kinds of Assessment Are Formative?

It follows from the proposed definition for formative assessment that any assessment that provides evidence that has the potential to improve instructional decision making can be formative, whether these decisions are taken by teachers, peers, or the learners themselves. The assessment might simply monitor the achievement of students, indicating that for some students, the instruction was unsuccessful. If the teacher then organizes additional instruction for those students, even if it is to go over the material again but more slowly, then this is potentially formative. If the assessment provides additional information that locates the precise nature of the students' difficulties, then it is diagnostic. The most useful assessments, however, are those that yield insights that are instructionally tractable. In other words, not only do they identify which students are having difficulties (the monitoring assessment) or locate the specific difficulties (the diagnostic assessment): They also yield insights into the kinds of next steps in instruction (including possibly steps to be taken by learners) that are likely to be most effective.

To give a concrete example, suppose a class has taken a test that assesses the ability to find the largest or smallest fraction in a given set. Knowing the scores of the students on this test would provide a monitoring assessment. It would identify those students who had mastered this skill sufficiently well to move on, and those who need more help. If the teacher organized additional instruction for these latter students, either by holding an additional class at the end of the day, or through the provision of targeted learning materials, the test would be formative (or more precisely, would function formatively), because the availability of the test scores allowed the teacher to make a better instructional decision than he or she would have been able to make in the absence of the information about the test scores.

If her test had been carefully constructed, there might also be *diagnostic* information in the students' responses. For example, the teacher might notice that most of the students who got low scores on the test had far greater success with items that included a number of unitary fractions (fractions with 1 as the numerator) than those without unitary fractions. Although this would be useful information, this insight does more to locate the learning difficulty than to indicate what should be done to overcome it—the teacher could focus instructional intervention on nonunitary fractions, which is likely to be more appropriate than reteaching the whole topic. However, if the teacher can see from the responses that many of the students are operating with a naïve strategy that the smallest fraction is the one with the largest denominator, and the largest fraction is the one with the smallest denominator—a strategy that is successful with unitary fractions (Vinner, 1997)—then this provides information for the teacher that is instructionally tractable. Such assessments not only signal the problem (monitoring) and locate it (diagnosing). They situate the problem within a theory of action that can suggest measures that could be taken to improve learning. The best formative assessment therefore identifies recipes for future action.

Note that in the three scenarios about the fraction item, in each case the assessment functioned formatively, because information was used to make instructional decisions that were likely to be better than those that would have been taken in the absence of the evidence. However, the fact that in all three cases the assessment functioned formatively did not mean that all three ways of using the evidence were likely to be equally effective. By definition, assessments that yield diagnostic insights are likely to lead to better instructional decisions than those that simply monitor student achievement, and those that yield insights that are instructionally tractable would be better still.

One of the differences between assessments that monitor, those that diagnose and those that yield insights that are instructionally tractable is a matter of the specificity of the information yielded—to be instructionally tractable, the assessment needs to yield more information than simply whether learning is taking place, or, if it is not, what specifically, is not being learned. But for an assessment outcome to be instructionally tractable, it must also entail theories of curriculum and theories of learning.

Instructional tractability entails a theory of curriculum because the focus is on answering the question: "What next?" This implies that there is a clear notion of a learning progression; that is, a description of the "knowledge, skills, understandings, attitudes or values that students develop in an area of learning, in the order in which they typically develop them" (Forster & Masters, 2004, p. 65). Instructional tractability

also entails a theory of learning, because before a decision can be made about what evidence to elicit, it is necessary to know not just what comes next in learning, but what kinds of difficulties learners have in making those next steps. The links between formative assessment and theories of learning are spelled out in more detail in Black and Wiliam (2005), Brookhart (2007), Wiliam (2007a), and Black and Wiliam (2009) and are summarized briefly in a subsequent section of this chapter, "A New Theory of Formative Assessment: Key Instructional Processes."

CYCLE LENGTHS FOR FORMATIVE ASSESSMENT

In the example of the fractions test discussed above, the action taken by the teacher follows quickly from the elicitation of the evidence about student achievement. In general, however, formative assessment allows for cycles of elicitation, interpretation, and action of any length, provided the information is used to inform instructional decisions. Consider the following six scenarios.

Scenario 1. In spring 2008, a science supervisor in a school district needed to plan the summer workshops that would be offered to eighth-grade science teachers in the district. She analyzed the scores obtained by the districts' eighth-grade students on the 2007 tests and noted that, whereas the average scores on science tests were comparable to the state average, performance on earth science items was much lower than the state average. The teacher decided to make earth sciences the focus of the professional development activities offered in summer 2008. The workshops were well attended by the district's eighth-grade science teachers. Teachers returned to school in fall 2008, and implemented revised instructional methods based on their learning over the summer. As a result, the achievement of eighth-grade students on earth sciences items improved in the tests taken in spring 2009.

Scenario 2. Each year, a group of high school teachers of Algebra I reviewed students' performance on a state-wide Algebra I test. They looked at the difficulty level (proportion correct) for each item on the test. Where item difficulties were lower than expected, they looked at how instruction on that aspect of the curriculum was planned and delivered, and at ways in which the instruction could be strengthened in the following year.

Scenario 3. A school district used a series of interim tests that were keyed to the curriculum and administered at intervals of 6 to 10 weeks to check on student progress. Students whose scores were below the threshold determined to be necessary to have an 80% chance of passing the state test were required to attend additional instruction on Saturdays.

Scenario 4. In elementary and middle school mathematics and science teaching in Japan, a teaching unit is typically allocated 13 or 14 lessons (Lewis, 2002). The content usually occupies only 10 or 11 of the lessons, allowing time for a short test to be given in the 11th or 12th lesson, and for the teacher to use the remaining lessons to reteach aspects of the unit that were not well understood.

Scenario 5. During the last 3 minutes of a lesson, a history teacher who had been

teaching about problems of bias in historical sources asked the students to answer, on a 3-inch by 5-inch index card, the question "Why are historians concerned about bias in historical sources?" The students turned in these "exit passes" as they left the class. The teacher read through the students' responses and then discarded the exit passes, having decided that the students' answers indicated a good enough understanding for the teacher to move on to a new chapter in the next lesson.

Scenario 6. A middle school science teacher had been teaching students to distinguish between different kinds of levers. After explaining that the key principle of the classification of levers concerns the relative arrangement of the load, the effort, and the fulcrum, she illustrated the principle with three examples: a see-saw (type 1), a wheel-barrow (type 2), and a deep sea fishing rod (type 3). To check on the students' understanding, she asked the class how a pair of tweezers would be classified, asking each student to hold up one, two, or three fingers to indicate their response. She was surprised that most of the students indicated that they thought the tweezers were a type 2 lever. When she asked them why, the students indicated that this was because there are two arms to the tweezers. She realized that it was necessary to introduce more examples, such as a pair of scissors and a nutcracker, because the students needed to understand that it is the relative distribution of the effort, load, and fulcrum that is important, not the number of components.

Now, let us recall the definition of formative assessment proposed by Black and Wiliam (2009):

> 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. (p. 6)

According to this definition, in each of the six scenarios, the assessment functioned formatively because evidence from the assessment was interpreted and used to make decisions that were likely to be better (or in the case of example 5, better founded) than the decisions that would have been made in the absence of that evidence. The length of the formative assessment cycle was also attuned to the capacity of the system to respond to the evidence generated—for example, there is little point in generating information on a daily basis if the decisions that the evidence is to inform are only taken on a monthly basis (Wiliam & Thompson, 2007).

However, many of these six scenarios would fail to be formative under some of the definitions discussed above. In particular, Shepard (2007) and Kahl (2005) might resist the idea that the use of assessment in examples 1, 2, and 3 were formative. They would likely point out that many test vendors have uncritically adopted the label *formative* and often have simply applied the label to tests originally designed to serve a summative function (see also Popham, 2006). Shepard (2007) argues that "what makes formative assessment *formative* is that it is immediately used to make adjustments so as *to form* new learning" (p. 281). Yet, in each of the six examples above, assessment evidence

Table 2.1 Cycle Lengths for Formative Assessment

Type	Focus	Length
Long-cycle	Across marking periods, quarters, semesters, years	4 weeks to 1 year
Medium-cycle	Within and between instructional units	1 to 4 weeks
Short-cycle	Within and between lessons	Day by day: 24 to 48 hours Minute by minute: 5 seconds to 2 hours

was used to make adjustments so as to form new learning. Examples 1, 2, and 3 fail to meet the requirement for immediacy imposed by Cowie and Bell (1999), Looney (2005), and Shepard (2007), but arguably, so also does example 4, depending on one's definition of immediacy.

The research literature supports the contention that the kinds of formative assessment illustrated in examples 4, 5, and 6 are more likely to increase learning, and by a greater amount, than the uses in examples 1, 2, and 3. Indeed, as Shepard (2007) argues, there is relatively little evidence that interventions such as examples 1, 2, and 3, are likely to have much impact at all. However, it seems odd to say that these examples are not formative in order to be able to reserve the term *formative* for those kinds of assessments that do make a significant difference to student outcomes. Rather, it would seem to make more sense—and to do less violence to the vernacular use of the word—to decide that where the assessment forms the direction of future learning, it can be described as formative, but to acknowledge that there are different kinds of cycle-length in formative assessment, as proposed by Wiliam and Thompson (2007), and shown in Table 2.1.

It is also, arguably, good *realpolitik* in that it seems unlikely that test publishers would agree to forgo the additional sales of their tests that they can expect from branding their tests as formative (and thus lay claim to a body of research about efficacy in practice) simply because they are asked to do so by researchers. The important question is therefore not, "Is this assessment formative?" but, "How does the use of this assessment improve learning?" and, echoing the conclusions of Kluger and DeNisi (1996), "How sustainably does this assessment improve learning?"

To answer this last question, and to understand what kinds of formative assessments are likely to be most effective, it is necessary to go beyond the functional definition of formative assessment, and look in more detail at the underlying processes.

A NEW THEORY OF FORMATIVE ASSESSMENT: KEY INSTRUCTIONAL PROCESSES

The systems approach to formative assessment proposed by Ramaprasad (1983), and which provides the basis for the definition of assessment for learning adopted by the Assessment Reform Group (Broadfoot et al., 2002), draws attention to three key instructional processes: (1) establishing where the learners are in their learning; (2) establishing where they are going; and (3) establishing what needs to be done to get them there.

The definition of formative assessment adopted here is based on a crossing of the

process dimension (where learners are in their learning, where they are going, how to get there) with that of the agent of the instructional process (teacher, peer, learner). The resulting nine cells can be collapsed into the five key strategies of formative assessment as shown in Figure 2.1 (Wiliam & Thompson, 2007). The focus of Figure 2.1 is the subject classroom. As Black and Wiliam (2005) observe, the activities that take place when students are learning mathematics are very different from those that take place when students are learning English language arts. The role of the students and the teacher, and the nature of their interactions with each other and with the discipline are likely to be different too. Furthermore, the subject classroom is, of course, nested within a school, which in turn is located in a community, and so on. Although it is beyond the scope of this chapter, any adequate account of formative assessment will have to acknowledge these multiple contexts. The stance taken in this chapter is that, ultimately, assessment must feed into actions in the subject classroom in order to affect learning; this simplification seems reasonable, at least as a first order approximation (see Black and Wiliam (2005) and Pryor and Crossouard (2005) for examples of sociocultural approaches to the implementation of formative assessment.

The framework represented by Figure 2.1 suggests that assessment for learning can be conceptualized as consisting of five key strategies (Wiliam & Thompson, 2007):

1. clarifying, sharing, and understanding learning intentions and criteria for success;
2. engineering effective classroom discussions, questions, and tasks that elicit evidence of learning;
3. providing feedback that moves learners forward;
4. activating students as instructional resources for one another; and
5. activating students as the owners of their own learning.

A detailed account of each of these five key strategies can be found in Wiliam (2007a). In the remainder of this chapter, each of the strategies is summarized briefly, and the

<table>
<tr><th></th><th>Where the learner is going</th><th>Where the learner is right now</th><th>How to get there</th></tr>
<tr><td>Teacher</td><td rowspan="3">Clarifying learning intentions and sharing and criteria for success (1)

Understanding and sharing learning intentions and criteria for success (1)

Understanding learning intentions and criteria for success (1)</td><td>Engineering effective classroom discussions. activities and tasks that elicit evidence of learning (2)</td><td>Providing feedback that moves learners forward (3)</td></tr>
<tr><td>Peer</td><td colspan="2">Activating students as instructional resources for one another (4)</td></tr>
<tr><td>Learner</td><td colspan="2">Activating students as the owners of their own learning (5)</td></tr>
</table>

Note: Numbers in parentheses indicate to which of the five key strategies an aspect relates

Figure 2.1 Aspects of formative assessment.

chapter concludes with some thoughts about future directions for research, theory, and practice.

Clarifying, Sharing, and Understanding Learning Intentions and Criteria for Success

The first strategy involves clarifying, communicating, and understanding learning intentions and criteria for success with students. At times it will be possible to specify the learning intentions in terms of clear goals, with narrowly drawn criteria for success; for example, when the teacher is trying to help students learn how to balance a chemical equation. At other times, particularly in creative work, such precision would be neither possible nor desirable, as when students are engaged in exploring the possibilities of painting with acrylics. In such situations, the teacher might be operating with a broad "horizon" (Black et al., 2003, p. 68) of possible, and acceptable, goals; different students can pursue different avenues. However, it is important to note that it is not the case that "anything goes." Although there may be a broad range of different directions in which learners might usefully go, there will be some that the teacher regards as unlikely to lead to useful learning, at which point the teacher would probably intervene to redirect the learner's activities.

An important consequence of this view of formative assessment is that, whereas it is necessary for there to be clarity about what is to be learned, what the learners are to learn is completely independent of formative assessment (Wiliam, 2007a). In other words, a commitment to formative assessment does not entail any particular view of what the learning intentions should be, nor does it entail a commitment to any particular view of what happens when learning takes place. This is important because, in many formulations of formative assessment, there is an implication that a commitment to formative assessment entails a commitment to certain kinds of learning goals; for example, to deep learning. While deep learning may indeed be desirable, it does not necessarily take place by a commitment to formative assessment, which can be used to help students reach instrumental or more shallow goals just as well as ultimate or deeper goals.

Even if learning intentions and criteria for success with students are clarified, communicated, and understood, it also makes no prescription about who determines the learning goal. While the youngest learners may have relatively little choice over what they are to learn, as they get older they will assume greater responsibility. However, even within further and higher education, where the student chooses courses of study, there will generally be an established curriculum, so that the actual learning intentions, and the associated success criteria, are likely to be a matter for negotiation between learner and teacher.

Engineering Effective Classroom Discussions, Activities, and Tasks that Elicit Evidence of Learning

The second strategy listed in Figure 2.1 focuses on the elicitation of evidence of achievement. While this elicitation will frequently take the form of questioning, it is important to note that any actions that elicit evidence that can be used to inform instruction are also included. For example, for teachers of students with multiple and profound learning

difficulties, it may be that evidence of learning is elicited by touch rather than through anything recognizable as a question.

The important point here is that not all elicited evidence is equally useful. Some kinds of evidence will support only a monitoring or a diagnostic function. As noted above, for the evidence elicited to be instructionally tractable, the evidence that is elicited and the way in which it is elicited will need to be driven by both a clear understanding of the learning intentions (whether defined narrowly or broadly) an understanding of progressions in learning (Heritage, 2008), and of the difficulties that learners experience.

However, it would be a mistake to assume that diagnostic assessments are always to be preferred to monitoring assessments, and those that yield instructionally tractable insights into learning are always to be preferred to diagnostic assessments because the range of available decisions might be limited. If the only available decision is whether to require the student to repeat the grade or not, then a simple assessment of the proportion of the intended learning that has been learned will be sufficient. A more diagnostic assessment would be required if the decision is "Which parts of this chapter do I need to review with the class before the end-of-chapter test?"

Nevertheless, in general, to be most effective, instruction needs to be tailored to the specific needs of individual learners, and so a greater range of instructional alternatives than simply repeating sequences of instruction will be required. For formative assessment to be instructionally tractable, the teacher must first be clear about the range of alternative instructional moves that are possible, should then decide what kinds of evidence would be useful in choosing among the relevant alternatives, and only then elicit the evidence needed to make the decision. In other words, the choice of what kind of evidence to elicit is driven by a theory of learning and almost all the intellectual heavy lifting is done before the teacher actually elicits the evidence of achievement.

Providing Feedback that Moves Learners Forward

The requirement for feedback that moves learning forward—the third strategy in Figure 2.1—emphasizes the fact that effective formative assessment is prospective, rather than retrospective. It is the view through the windshield rather than the rear-view mirror or, as Douglas Reeves has memorably suggested, it is the difference between a medical examination and a postmortem (personal communication, October 31, 2008). This encapsulates the two key findings of Kluger and DeNisi (1996) and Hattie and Timperley (2007) discussed above: (1) that it is more productive to think about the processes that are triggered by the feedback intervention, and (2) that feedback interventions are likely to be more effective if they cue attention to the task, how the learner works on the task, and the processes of self-regulation in which the learner engages rather than cue attention to the self. Perhaps even more simply, feedback is likely to be more effective when it causes a cognitive rather than an affective reaction. Of course, whether this happens depends not only on the quality of the feedback, but also on the learner, and the learning milieu in which the feedback is given and received (Black & Wiliam, 2005, 2009)

The other aspect of feedback that moves learning forward is related to instructional adjustments. Instead of providing feedback to the learner, the assessment outcomes may instead provide feedback for the teacher so that he or she can modify the instruction in order to be more effective (whether for the students on whom the data were collected

or some other students being taught at some point in the future). In other words, the assessment might be more formative for the teacher than the student.

Activating Students as Owners of Their Own Learning

The last two of the key strategies listed in Figure 2.1 are related to the role of learners in the formative assessment process, including the extent to which students are owners of their own learning and active as learning resources for each other and, for convenience, are here discussed in the reverse order of their appearance in Figure 2.1. For students to become owners of their own learning they need both to own the curricular objectives, and to be active in guiding their own learning—in other words, they must become self-regulated learners. The notion of self-regulated learning is a rich focus of inquiry, with a vast literature of its own, most of which is highly relevant to the notion of formative assessment. Below, a brief summary of some of the most important points is presented so that the interested reader can pursue them in more detail.

Winne (1996) defined self-regulated learning as a "metacognitively governed behavior wherein learners adaptively regulate their use of cognitive tactics and strategies in tasks" (p. 327). Others have pointed out that learners often possess, but do not deploy, the necessary self-regulation skills, and that the problem may be a lack of motivation or volition (Corno, 2001). Still others have argued for the need to look at issues of self-regulation with broader theoretical frames including sociocultural (Hickey & McCaslin, 2001; McCaslin & Hickey, 2001) or social constructivist (Op't Eynde, DeCorte, & Verschaffel, 2001) perspectives.

One of the most general definitions of self-regulation is provided by Boekaerts (2006), who defines the concept as "a multilevel, multicomponent process that targets affect, cognitions, and actions, as well as features of the environment for modulation in the service of one's goals" (p. 347). According to Boekaerts, distinguishing between cognitive and motivational aspects of self-regulated learning is difficult because self-regulated learning is both metacognitively governed *and* affectively charged.

A number of ways of bringing together the motivational and cognitive perspectives on self-regulation have been proposed; summaries of some of these can be found in Wiliam (2007a). For the purpose of this chapter, and in particular in terms of the strategy of activating students as owners of their own learning, a model that is particularly relevant is the *dual processing theory* developed by Boekaerts (1993). According to Boekaerts:

> It is assumed that students who are invited to participate in a learning activity use three sources of information to form a mental representation of the task-in-context and to appraise it: (1) current perceptions of the task and the physical, social, and instructional context within which it is embedded; (2) activated domain-specific knowledge and (meta)cognitive strategies related to the task; and (3) motivational beliefs, including domain-specific capacity, interest and effort beliefs. (2006, p. 349)

When the task appraisal is positive, energy is activated along the growth pathway where the goal is to increase competence. Boekaerts describes this sort of self-regulation

as *top-down* because the flow of energy is directed by the student. Attention shifts toward the well-being pathway, where the goal is to prevent threat, harm, or loss when the task appraisal is negative. This form of self-regulation is termed *bottom-up* by Boekaerts because it is triggered by cues in the environment, rather than by learning goals. Where such bottom-up regulation is the norm, then learning is obviously compromised. However, in certain cases it can be positive because, by temporarily attending to well-being, the student may find a way to shift energy and attention back to the growth pathway.

Of course, the relationship between top-down and bottom-up pathways of regulation is dynamic, rather than being a stable feature of an individual learner. Boekaerts (2001) found no direct link between domain-specific motivational beliefs and learning intention in any of the mathematics classrooms under study; students' decisions about whether to invest effort in a mathematics assignment depended primarily on their appraisal of the specific task in front of them, although Ross, Rolheiser, and Hogaboam-Gray (2002) found that students' decisions about whether to invest effort were also influenced by friends and parents.

One of the major strengths of the dual-processing model is that it supports the integration of a variety of different perspectives on the broad idea of activating students as owners of their own learning, including the relationship between motivation and interest, the way that learners attribute their successes and failures in learning, and the way they develop ideas about their self-efficacy.

For example, when students are interested in a task, they are likely to engage in activity along the growth pathway (Hidi & Harackiewicz, 2000). When students are not personally interested in a task, interest may be sparked by something in the task situation, thus also triggering activity along the growth pathway. Where interest is not the main driver of attention, considerations of task value versus cost will become important (Eccles et al., 1983). In terms of the theories of motivation proposed by Deci and Ryan (1994), activity along the growth pathway is associated with motivation stemming from values within the individual while activity along the well-being pathway is associated with values originating outside the individual. In terms of achievement goal theory (Dweck & Leggett, 1986), students displaying mastery orientation are likely to be activating the growth pathway, while those displaying performance orientation are likely to be activating the well-being pathway.

Self-efficacy beliefs (Bandura, 1977) can drive progress along either pathway. Along the growth pathway, self-efficacy drives adaptive cognitive and metacognitive strategy use, whereas along the well-being pathway, self-efficacy beliefs are likely to steer the learner away from performance-avoidance goals and toward performance-approach goals. Similarly views of ability as incremental (Dweck, 2000) help the learner stay on the growth pathway, whereas entity views of ability direct activity toward the well-being pathway, where details of the task-in-context, appraised in the light of views of personal capability, will influence decisions about whether to engage in the task.

Activating Students as Learning Resources for One Another

The final strategy listed in Figure 2.1 is to activate students as learning resources for one another. In some ways this strategy provides a focus for the other four strategies,

because it combines aspects of each of them. In order for students to assess the work of others, they have to internalize the learning intentions or the success criteria, and these understandings then become available to the students for use in their own productions (Black et al., 2003). Furthermore, because assessing someone else's work is less emotionally charged than attempting to assess one's own, peer-assessment provides a useful stepping-stone to effective self-assessment, and thus to improved self-regulation in learning (Black et al., 2003, p. 62). In peer tutoring and in other forms of collaborative learning, the peer is frequently cast in the role of teacher, so eliciting evidence and providing feedback are foremost. Indeed, the boundaries between the strategies frequently become blurred. When teachers ask students to review their learning by constructing test items (with correct answers) as studied by Foos, Mora, and Tkacz (1994) students need to think carefully about the learning intentions of the work they have been studying, and about what makes a good way of eliciting evidence. When such items are administered to other learners (Fontana & Fernandes, 1994), students are active as learning resources for one another, and are therefore also improving their own skills of self-regulation.

SUMMARY AND SUGGESTIONS FOR FUTURE WORK

This chapter has provided a brief history of the idea of formative assessment, together with a review of the research that supports its efficacy in educational settings. While there are inevitable methodological problems in synthesizing the results from studies that use different instruments to measure outcomes and are conducted in different traditions, there can be little doubt that increased use of formative assessment is one of the most educationally effective and most cost effective ways of increasing student achievement. Moreover, the effects appear to be generalizable across learning of different types, in a range of contexts, and for learners of all ages.

As the idea of formative assessment has developed, the definition of the term *formative* has ranged from a description of the timing of an assessment (any assessment before "the big one") to a description of a kind of instrument. However, since the evidence from an assessment instrument can be used in a range of ways, this chapter has proposed a definition of formative assessment in terms of the extent to which evidence of learner achievement is used to inform decisions about teaching and learning. In particular, formative assessment is concerned with the creation of, and capitalization upon, moments of contingency in instruction (including both teaching and learning) with a view to regulating learning processes more effectively.

Although somewhat abstract in its formulation, this definition supports immediate application to educational settings in terms of five key strategies:

1. clarifying, sharing and understanding learning intentions and criteria for success;
2. engineering effective classroom discussions, questions, and tasks that elicit evidence of learning;
3. providing feedback that moves learners forward;
4. activating students as the owners of their own learning; and
5. activating students as instructional resources for one another.

The five strategies are, of course, not the only important processes in instruction, but they do appear to be powerful lenses for thinking about practice, and thus for supporting teachers in engaging with wider issues of psychology, pedagogy, and curriculum.

As Kluger and DeNisi (1996) have suggested, further studies designed to identify more precisely the size of impact on student learning that can be achieved with formative assessment are unlikely to be helpful. What is likely to be helpful are studies that relate the kinds of feedback interventions to the learning processes they engender. Such studies, conducted over extended periods of time (at least a year) would also show whether high quality instruction is compatible with increased success on standardized tests, which will be important in developing an understanding of how to improve instruction in settings that make extensive use of tests that are used to hold students and teachers accountable. Without such evidence, attempts at reform are likely to be met with the reactions such as: "I'd love to teach for deep understanding, but I have to raise my test scores."

However, such studies are likely to be ultimately far less important than studies of how to support teachers in making greater use of formative assessment in their own practice. Certainly, everything about what makes for the most effective uses of formative assessment has not yet been discovered; however, enough is known to build a substantial consensus around the kinds of classrooms that are most effective. Far less is known about how to get more such classrooms. As Black and Wiliam (1998a) pointed out:

> It is hard to see how any innovation in formative assessment can be treated as a marginal change in classroom work. All such work involves some degree of feedback between those taught and the teacher, and this is entailed in the quality of their interactions which is at the heart of pedagogy. (p. 16)

There are some success stories here (e.g., Wiliam, Lee, Harrison, & Black, 2004), but very little is known about the factors that support the implementation of educational innovations at scale (Coburn, 2003; Thompson & Wiliam, 2008). In order to secure the improvements in educational outcomes that the existing research on formative assessment has shown is possible, designing ways of supporting teachers to develop their practice of formative assessment at scale must be the main priority.

REFERENCES

Alexander, R. (2008). *Essays on pedagogy*. York, UK: Dialogos.

Allal, L., & Lopez, L. M. (2005). Formative assessment of learning: A review of publications in French. In J. Looney (Ed.), *Formative assessment: Improving learning in secondary classrooms* (pp. 241–264). Paris: Organization for Economic Cooperation and Development.

Bandura, A. (1977). Self-efficacy: Towards a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Buckingham, UK: Open University Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 8–21.

Black, P. J., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7–73.

Black, P. J., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.
Black, P., & Wiliam, D. (2005). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 81–100). London: Sage.
Black, P. J., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability, 21*(1), 5–31.
Bloom, B. S. (1984). The search for methods of instruction as effective as one-to-one tutoring. *Educational Leadership, 41*(8), 4–17.
Boekaerts, M. (1993). Being concerned with well being and with learning. *Educational Psychologist, 28*(2), 149–167.
Boekaerts, M. (2001). Context sensitivity: Activated motivational beliefs, current concerns and emotional arousal. In S. Volet & S. Järvelä (Eds.), *Motivation in learning contexts: Theoretical advances and methodological implications* (pp. 17–31). Oxford, England: Pergamon.
Boekaerts, M. (2006). Self-regulation and effort investment. In K. A. Renninger & I. E. Sigel (Eds.), *Handbook of child psychology: Vol. 4. Child psychology in practice* (6th ed., pp. 345–377). New York: Wiley.
Broadfoot, P. M., Daugherty, R., Gardner, J., Gipps, C. V., Harlen, W., James, M., et al. (1999). *Assessment for learning: Beyond the black box*. Cambridge, UK: University of Cambridge School of Education.
Broadfoot, P. M., Daugherty, R., Gardner, J., Harlen, W., James, M., & Stobart, G. (2002). *Assessment for learning: 10 principles*. Cambridge, UK: University of Cambridge School of Education.
Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record, 106*(3), 429–458.
Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 43–62). New York: Teachers College Press.
Coburn, C. (2003). Rethinking scale: moving beyond numbers to deep and lasting change. *Educational Researcher, 32*(6), 3–12.
Corno, L. (2001). Volitional aspects of self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated leaning and academic achievement: Theoretical perspectives* (2nd ed., pp. 191–225). Hillsdale, NJ: Erlbaum.
Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education: Principles, Policy, and Practice, 6*(1), 32–42.
Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438–481.
Deci, E. L., & Ryan, R. M. (1994). Promoting self-determined education. *Scandinavian Journal of Educational Research, 38*(1), 3–14.
Dempster, F. N. (1991). Synthesis of research on reviews and tests. *Educational Leadership, 48*(7), 71–76.
Dempster, F. N. (1992). Using tests to promote learning: A neglected classroom resource. *Journal of Research and Development in Education, 25*(4), 213–217.
Denvir, B., & Brown, M. L. (1986a). Understanding of number concepts in low-attaining 7-9 year olds: Part 1. Development of descriptive framework and diagnostic instrument. *Educational Studies in Mathematics, 17*(1), 15–36.
Denvir, B., & Brown, M. L. (1986b). Understanding of number concepts in low-attaining 7–9 year olds: Part II. The teaching studies. *Educational Studies in Mathematics, 17*(2), 143–164.
Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality and development*. Philadelphia: Psychology Press.
Dweck, C. S., & Leggett, E. L. (1986). Motivational processes affecting learning. *American Psychologist, 41(*10), 1040–1048.
Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., et al. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco: W. H. Freeman.
Elshout-Mohr, M. (1994). Feedback in self-instruction. *European Education, 26*(2), 58–73.
Fontana, D., & Fernandes, M. (1994). Improvements in mathematics performance as a consequence of self-assessment in Portugese primary school pupils. *British Journal of Educational Psychology, 64*(4), 407–417.
Foos, P. W., Mora, J., & Tkacz, S. (1994). Student study techniques and the generation effect. *Journal of Educational Psychology, 86*(4), 567–576.
Forster, M., & Masters, G. N. (2004). Bridging the conceptual gap between classroom assessment and accountability. In M. Wilson (Ed.), *Towards coherence between classroom assessment and system accountability:*

103rd Yearbook of the National Society for the Study of Education (Part II, pp. 51–73). Chicago: University of Chicago Press.

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

Gipps, C. V., & Stobart, G. (1997). *Assessment: A teacher's guide to the issues* (3rd ed.). London: Hodder and Stoughton.

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

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: Council of Chief State School Officers.

Hickey, D. T., & McCaslin, M. (2001). A comparative, sociocultural analysis of context and motivation. In S. Volet & S. Järvelä (Eds.), *Motivation in learning contexts* (pp. 33–55). Oxford, UK: Pergamon.

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70*(2), 151–179.

James, M. (1992, April). *Assessment for learning.* Assembly session at the annual conference of the Association for Supervision and Curriculum Development, New Orleans, LA.

Kahl, S. (2005, September 21). Where in the world are formative tests? Right under your nose! *Education Week, 25*(4), 11.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.

Köller, O. (2005). Formative assessment in classrooms: A review of the empirical German literature. In J. Looney (Ed.), *Formative assessment: Improving learning in secondary classrooms* (pp. 265–279). Paris: Organization for Economic Cooperation and Development.

Lewis, C. C. (2002). *Lesson study: A handbook of teacher-led instructional change*. Philadelphia: Research for Better Schools.

Looney, J. (Ed.). (2005). *Formative assessment: Improving learning in secondary classrooms*. Paris: Organisation for Economic Cooperation and Development.

McCaslin, M., & Hickey, D. T. (2001). Educational psychology, social constructivism, and educational practice: A case of emergent identity. *Educational Psychologist, 36*(2), 133–140.

Mitchell, R. (1992). *Testing for learning*. New York: Free Press.

National Assessment of Educational Progress. (2006). *The Nation's Report Card: Mathematics 2005* (Vol. NCES 2006-453). Washington, DC: Institute of Education Sciences.

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

Nyquist, J. B. (2003). *The benefits of reconstruing feedback as a larger system of formative assessment: A meta-analysis.* Unpublished master's thesis. Nashville, TN, Vanderbilt University.

Op't Eynde, P., DeCorte, E., & Verschaffel, L. (2001). "What to learn from what we feel?" The role of students' emotions in the mathematics classroom. In S. Volet & S. Järvelä (Eds.), *Motivation in learning contexts: Theoretical advances and methodological implications* (pp. 149–167). Oxford, UK: Pergamon.

Popham, W. J. (2006). Phony formative assessments: Buyer beware! *Educational Leadership, 64*(3), 86–87.

Popham, W. J. (2007, April). *Determining the instructional sensitivity of accountability tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Pryor, J., & Crossouard, B. (2005, September). *A sociocultural theorization of formative assessment.* Paper presented at Sociocultural Theory in Educational Research and Practice Conference, Brighton, UK.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioural Science, 28*(1), 4–13.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (2002). Influences on student cognitions about evaluation. *Assessment in Education: Principles, Policy, and Practice, 9*(1), 81–95.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369–393.

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

Shepard, L. A. (2007). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279–303). Mahwah, NJ: Erlbaum.

Shepard, L. A., Hammerness, K., Darling-Hammond, L., Rust, F., Snowden, J. B., Gordon, E., et al. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 275–326). San Francisco, CA: Jossey-Bass.

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

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758–765.

Sutton, R. (1995). *Assessment for learning*. Salford,UK: RS Publications.

Thompson, M., & Wiliam, D. (2008). Tight but loose: A conceptual framework for scaling up school reforms. In E. C. Wylie (Ed.), *Tight but loose: Scaling up teacher professional development in diverse contexts* (RR-08-29, pp. 1–44). Princeton, NJ: Educational Testing Service.

Vinner, S. (1997). From intuition to inhibition: Mathematics, education and other endangered species. In E. Pehkonen (Ed.), *Proceedings of the 21st conference of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 63–78). Lahti, Finland: University of Helsinki Lahti Research and Training Centre.

Wiener, N. (1948). *Cybernetics, or the control and communication in the animal and the machine*. New York: Wiley.

Wiliam, D. (2007a). Keeping learning on track: Classroom assessment and the regulation of learning. In F. K. Lester Jr. (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053–1098). Greenwich, CT: Information Age.

Wiliam, D. (2007b, April). *An index of sensitivity to instruction.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Wiliam, D. (2007c). Content *then* process: Teacher learning communities in the service of formative assessment. In D. B. Reeves (Ed.), *Ahead of the curve: The power of assessment to transform teaching and learning* (pp. 183–204). Bloomington, IN: Solution Tree.

Wiliam, D. (2008). International comparisons and sensitivity to instruction. *Assessment in Education: Principles, Policy, and Practice, 15*(3), 253–257.

Wiliam, D., & Black, P. J. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22*(5), 537–548.

Wiliam, D., Lee, C., Harrison, C., & Black, P. J. (2004). Teachers developing assessment for learning: impact on student achievement. *Assessment in Education: Principles Policy and Practice, 11*(1), 49–65.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Mahwah, NJ: Erlbaum.

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences, 8*, 327–353.