FORMATIVE ASSESSMENT: DEFINITIONS AND RELATIONSHIPS

Dylan Wiliam, Institute of Education, University of London

Abstract

The idea that assessment is intrinsic to effective instruction is traced from early experiments in the individualization of learning through the work of Benjamin Bloom to reviews of the impact of feedback on learners in classrooms. While many of these reviews detailed the adverse impact of assessment on learning, they also indicated that under certain conditions assessment had considerable potential to enhance learning. It is shown that understanding the impact that assessment has on learning requires a broader focus than the feedback intervention itself, particularly the learner's responses to the feedback, and the learning milieu in which the feedback operates. Different definitions of the terms "formative assessment" and "assessment for learning" are discussed, and subsumed within a broad definition that focuses on the extent to which instructional decisions are supported by evidence. The paper concludes by exploring some of the consequences of this definition for classroom practice.

Introduction

Almost half a century ago, David Ausubel suggested that the most important factor influencing learning is what the learner already knows, that teachers should ascertain this, and teach accordingly (Ausubel, 1968).

At the time, and perhaps even now, such a prescription might seem simple, but students do not learn what they are taught. Even when instruction is planned with great care, delivered effectively, and in a way that engages students, the learning outcomes often bear little or no relation to what was intended. If what a student learns as a result of a particular sequence of instructional activities is impossible to predict, even in the unlikely event that all the learners in an instructional group are at the same place when the instruction starts, within minutes, students will have reached different understandings. That is why assessment is a, perhaps the, central process in effective instruction. It is only through assessment that we can find out whether a particular sequence of instructional activities has resulted in the intended learning outcomes.

For many years, the word "assessment" was used primarily to describe processes of evaluating the effectiveness of sequences of instructional activities when the sequence was completed. The actions that guided learning processes before the end of the sequence were generally not regarded as kinds of assessment. Within the French language literature, they were typically discussed as aspects of the regulation of learning processes, and within the English language literature, to the extent that it was discussed at all, as simply an aspect of good teaching. More recently, particularly in the English-speaking research community, however, there has been an increasing tendency to seek to understand activities that are intended to guide the learning towards the intended goal, and that take place during the learning process, as forms of assessment. In this paper I review this development, and attempt to clarify the meanings of the terms *formative assessment* and *assessment for learning*.

Assessment for learning: origins and antecedents

For many years, it seems that the prevailing view of education was that, provided instruction was of reasonable quality, it need not be adaptive to the needs of learners. It was assumed either that well-designed instruction would be effective for the majority of students for whom it was intended (with others being assigned to remedial activities) or that the causes of any failures to learn lay within the individual learner (the material was just too hard for them, and they should instead pursue other, and generally less academic, avenues). However, in the 1960s, Benjamin Bloom and his graduate students at the University of Chicago began to explore the idea that the normal distribution of student outcomes was not a "natural" outcome, but caused by the failure of the instruction to recognize differences in learners.

The "Individual System" often regarded as the first truly individualized system of instruction (see, for example, Reiser, 1986), was developed by Frederic Burk, from 1912 to 1913, for use in the elementary school associated with the San Francisco Normal State School, an institution providing pre-service education for teachers. One of Burk's colleagues, Mary Ward, had been getting her trainee teachers to develop self-instructional materials and Burk and others developed similar materials that covered most of the curriculum from kindergarten to sixth grade. Two other individuals who had worked with Ward and Burk at the San Francisco Normal State School, Carleton Washburne and Helen Parkhurst, developed these ideas further after they left the School. In 1919, Washburne implemented the Winnetka Plan, when he became superintendent of the Winnetka Public Schools in Illinois and in the same year, Parkhurst implemented the Dalton Plan in a school for disabled students in Dalton, Massachusetts (Parkhurst, 1922).

Bloom was convinced that such individualization was beneficial—indeed he regarded one-to-one tutorial instruction as the "gold standard" for education against which others should be compared (Bloom, 1984a)—but was concerned that this obviously would not be affordable for mass public education, hence "the search for methods of group instruction as effective as one-to-one tutoring" (Bloom, 1984b).

One of the main reasons that one-to-one tutoring is so effective, according to Bloom, is that the tutor is able to identify errors in the student's work immediately, and then to provide clarification, and further follow-up if necessary (Guskey, 2010). Bloom described these two processes as "feedback" and "correctives" and this language has become part of the standard way of talking about assessment ever since. However, in a very important sense, Bloom's distinction between "feedback" and "correctives" has been counterproductive, and has served to distort the original meaning of the term "feedback" in a particularly unfortunate manner.

In 1940, Norbert Wiener and his colleagues had been developing automatic range-finders for anti-aircraft guns. He realized that effective action required a closed system that allowed the effects of actions taken within the system to be evaluated, and in the light of that evaluation, to modify future actions (Wiener, 1948). In such systems, there were two kinds of loops: those that tended to push the system further in the direction in which it was already going (which he termed positive feedback loops) and those that opposed the tendency in the system (which he termed negative feedback loops). Positive feedback loops produce instability, driving the system towards either explosion or collapse. Examples of the former are simple population growth with plentiful food and no predators, and inflationary price/wage spirals in economics; examples of the latter include economic depression, food hoarding in times of shortage, and the loss of tax revenue in urban areas as a result of "middle-class flight". The use of the qualifier "positive" is not intended to provide any assessment of the value of such feedback—indeed, positive feedback is almost always has undesirable effects. Rather the term "positive" denotes simply the alignment between the existing tendency of the system, and the effect of the impetus provided by the feedback.

In contrast, negative feedback loops produce stability, because they tend to drive the system back to a former state. One example of such a system is population growth with limited food supply, in which shortage of food slows population growth. Depending on the conditions, the system then either approaches, or oscillates with decreasing amplitude around, a steady state (the carrying capacity of the environment). Perhaps the most familiar example is the domestic thermostat. When the temperature of the room drops below the setting on the thermostat, a signal is sent to turn on the heating system. When the room heats up above the setting on the thermostat, a signal is sent to turn off the heating system.

The important point about Wiener's formulation is that information does not become "feedback" unless it is provided within a system that can use that information to affect future performance. The importance of thinking about feedback *systems*, rather than just the nature of the information itself, particularly within the behavioural sciences, was emphasized by Ramaprasad (1983) who noted: "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" (Ramaprasad, 1983, p. 4). The *use* of the information was reinforced by Sadler (1989):

> 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)

This is why Bloom's formulation is unhelpful. In describing the information generated about the gap between current and desired performance as "feedback" Bloom separated the information from its instructional consequences. For Wiener, Ramaprasad, and Sadler, feedback is more than just information. It is information generated within a particular system, for a particular purpose. This is why feedback has to be domain-specific. In helping players improve their free-throw percentage, basketball coaches don't just tell the athletes to make sure that they get the ball through the hoop; they focus on mechanics such as reminding the athlete to bend their knees, and to keep the elbows tucked in. When providing feedback to students learning mathematics, it is not helpful to tell them that they need to improve their work, even if this is true. It is more helpful to point out what kinds of errors they are making, and what they need to do to improve.

A second point about the feedback system is that it has been designed so that the information generated is capable of affecting the future performance of the system. As soon as the term "feedback" is used to denote simply any data about the gap between current and desired level of performance, or worse, simply for descriptions of the current level of performance, it loses all connection with its original, and powerful, meaning.

This is more than a semantic point. Originally, the suffix “back” was intended to describe the direction of information flow. In much current usage, it appears to be used to signify chronology, and it seems that any information about previous performance would qualify as feedback. Ramaprasad and Sadler suggest that the term “feedback” should not be used unless there is an impact on future performance. Others (e.g., Black & Wiliam, 1998b) accept the term “feedback” as it is commonly used, but require an additional condition, that it actually improves student learning, for it to be counted as good. Either way, what is important is the acknowledgement that the use of assessment information to improve learning cannot be separated from the instructional system within which it is provided. Between 1986 and 1998, nine substantial articles reviewed the impact of assessment practices on students and their learning in the context of the classroom, and the consistency of findings from these nine reviews created significant interest amongst researchers, policy-makers and classroom practitioners in how these ideas could be applied in practice. In the following section, these nine reviews, and more recent work in this area, are briefly reviewed.

Reviews of Research on Assessment and Classroom Learning

Fuchs and Fuchs (1986) conducted a meta-analysis of 21 research reports, which yielded 96 different effect sizes, relating to learners from pre-school to twelfth grade, most of whom had mild or moderate learning disabilities. All of the studies focused on the use of the feedback to and by teachers, with frequencies of between two and five times per week. The mean effect size was 0.70 standard deviations with slightly smaller effects (0.63) for the 22 effect sizes involving learners without disabilities. In those studies where teachers worked to set rules about reviews of the data and actions to follow (about half of the studies reviewed), the mean effect size was 0.92, whereas when actions were left to teachers’ judgments the effect size was only 0.42. Also noteworthy was that where teachers undertook to produce graphs of the progress of individual children as a guide and stimulus to action, the mean effect was larger (0.70) than in those where this was not done (mean effect size 0.26).

Another review (Natriello, 1987) proposed a model of the evaluation process as consisting of eight stages:

1. Establishing the purpose of the evaluation;

2. Assigning tasks to students;

3. Setting criteria for student performance;

4. Settings standards for student performance;

5. Sampling information on student performance;

6. Appraising student performance;

7. Providing feedback to student performers; and

8. Monitoring outcomes of the evaluation of students.

His major conclusion was that little could be concluded from the research, for three main reasons. The first was that the available studies researched what was actually in place, rather than what might be, and as a result tended to confirm the problems with existing practices rather than providing a basis for improved practice. The second was the fact that most of the studies reviewed "concentrate on one or two aspects of the evaluation process. As a result, they fail to consider the impact of other key elements in determining the effects of evaluations" (p. 170).

Third, and perhaps most importantly, few of the studies Natriello reviewed considered explicitly the fact that evaluations were used in schools for a multiplicity of purposes and that comparisons were likely to be misleading where different kinds of evaluations were compared in terms of functions for which they had not been designed. For example, a finding that differentiated feedback had more impact on directing future student learning than grades may be showing nothing more than the fact that systems generally do more effectively those things they are designed to do than those things they are not designed to do.

The third review (Crooks, 1988) had a rather narrower focus—the impact of classroom evaluation practices on students. His review covered formal classroom-based assessments such as tests, informal evaluation processes such as adjunct questions in texts, and oral questioning by teachers in class. His main conclusion was that "Too much emphasis has been placed on the grading function of evaluation and too little on its role in assisting students to learn" (p. 468). He also noted that a rebalancing of the attention paid to these two roles was needed, since an over-emphasis on the grading function not only used time that could more profitably spent on other activities, but was actually counter-productive, resulting in:

> reduction of intrinsic motivation, debilitating evaluation anxiety, ability attributions for success and failure that undermine student effort, lowered self-efficacy for learning in the weaker students, reduced use and effectiveness of feedback to improve learning, and poorer social relationships among the students. (p. 468)

A fourth study of the impact of classroom assessment practices on students and their learning in the context of the classroom was undertaken by Bangert-Drowns, Kulik and Kulik (1991), who investigated the effects of frequent classroom testing. They found that students who took at least one test over a 15 week period scored 0.5 standard deviations higher than those who did not, and that more frequent testing was associated with higher levels of achievement, although testing more frequently that once every two weeks appeared to confer no additional benefit. In a related study, Bangert-Drowns, Kulik, Kulik and Morgan (1991) reported the results of a meta-analysis of 58 effect sizes from 40 research reports on the effects of feedback in what they called "test-like" events such as questions embedded in programmed learning materials, review tests at the end of a block of teaching, and so on. They found that the crucial mediating variable in determining the impact of feedback on learning was the degree to which the nature of the feedback, and the way it was provided, encouraged "mindfulness" in students. So, where students could look ahead to the answers before they had attempted the questions themselves, achievement was reduced, but where the studies controlled for this "pre-search availability," the effect size was 0.26 standard deviations. Moreover, where the interventions also provided feedback to students in terms of details of the correct answer, rather than just whether the students' response was correct or not, the mean effect size was 0.58 standard deviations.

These broad findings were confirmed by Dempster (1991) in a review of studies on the effects of feedback in tests, although he pointed out that many of the relevant studies measured achievement in terms of content knowledge and low-level skills, so it was not clear that such findings would necessarily generalize to higher-order thinking. In a subsequent paper (Dempster, 1992), he argued that, while the benefits of integrating assessment with instruction were clear and there was an emerging consensus in the research for the conditions for effective assessment, including frequent testing soon after instruction, cumulating demand, and feedback soon after testing, assessment was neglected in teacher education and there was evidence that current practices in schools were far from these ideals.

A review by Elshout-Mohr (1994), published originally in Dutch, and reviewing many studies not available in English, suggested that knowledge of correct answers was more useful for simple tasks but less useful for more complex tasks. Correcting what is wrong can be an effective approach for the learning of simple material, but for more complex material, learning requires the development of new capabilities that, according to Elshout-Mohr, requires a more dialogic kind of feedback, rather than the giving of correct answers, and therefore also requires the learner to become active in managing the process.

Much of this work had focused on the effects of feedback in schools. However, in 1996, Kluger and DeNisi published a review of the effects of feedback in schools, colleges and workplaces. They began by defining feedback interventions as "actions taken by (an) external agent(s) to provide information regarding some aspect(s) of one's task performance" (p. 255). They identified over 3000 relevant studies published between 1905 and 1995, but noted that many of these were very small studies (in many cases involving only a single participant), and were reported in insufficient detail to permit the calculation of an effect size for the intervention. In order to be sure that poor quality studies were not being included, Kluger and DeNisi established three criteria for inclusion in their review:

1. The participants had to be divided into two groups, the only difference between the groups, as far as could be judged, being whether they received feedback or not.

2. The study involved at least ten participants.

3. They included a measurement of performance with sufficient details provided for the size of the impact of feedback on performance to be calculated.

Surprisingly, only 131 of the 3000 relevant studies satisfied these criteria, and these selected studies reported 607 effect sizes, involving 23,663 observations of 12,652 participants. Across all the studies, the average effect size for feedback was 0.41 standard deviations, but the effects varied considerably across the different studies. Most notably, in 50 out of the 131 studies (i.e., 38%), feedback actually lowered average performance. In seeking to understand this, they looked for "moderators" of feedback effects and found that feedback interventions were least effective when they focused attention on the self, more effective when they focused on the focal task, and most effective when they focused on the details of the focal task and when they involved goal-setting.

However, they concluded that whether feedback "works" or not, and if so, by how much, were not the right questions to ask. They pointed out that there are two kinds of feedback intervention: those that indicate that current performance falls short of the desired goal or that current performance exceeds the current goal. Once received, there are four kinds of response the individual can make—change behaviour to reach the goal, modify the goal, abandon the goal, or reject the feedback. This leads to eight possible effects of feedback interventions, as shown in table 1.

Table 1: Possible Responses to Feedback Interventions (Kluger & DeNisi, 1996)

Response type	Feedback indicates performance exceeds goal	Feedback indicates performance falls short of goal
Change behavior	Exert less effort	**Increase effort**
Change goal	**Increase aspiration**	Reduce aspiration
Abandon goal	Decide goal is too easy	Decide goal is too hard
Reject feedback	Feedback is ignored	Feedback is ignored

In other words, there are eight possible responses to a feedback intervention, and six of them are likely to be ineffective or worse. Only two responses, highlighted in bold in table 1, are likely to have positive outcomes. But even when feedback is seen to have an effect, this may not be sustained, as has been found in some implementations of computer-assisted learning. If feedback works to increase motivation, then increasingly large efforts need to be made to maintain motivation. Feedback focused on task learning can emphasize instrumental goals, and thus inhibit deep learning. In such situations it might be better to provide more task information or even to encourage a trial and improvement strategy, thus generating feedback without a feedback intervention. They showed that feedback interventions were less effective when they cued attention beyond the task (for example on the self), and more effective when the feedback cued attention on task motivation or task learning (this is taken up in more detail in the discussion of the review by Hattie and Timperley and of the "dual-processing" theory of Monique Boekaerts below). This model accounts for the well-known earlier findings that praise (like other cues that draw attention away from the task and towards the self) often has negative effects (Brophy, 1981).

Black and Wiliam (1998a) sought to update the reviews discussed above. One problem they reported was a general difficulty in defining the field. They noted that the reviews by Natriello and Crooks mentioned above had cited 91 and 241 references respectively, and yet only nine references were common to both papers and neither cited the review by Fuchs and Fuchs. In their own work, Black and Wiliam found that attempting to rely on electronic methods either generated far too many irrelevant sources or failed to identify key papers. In order to be sure of reviewing the field thoroughly, they physically examined each issue of 76 of the journals considered most likely to contain relevant research published between 1987 and 1997. Black and Wiliam's review, which cited 250 studies, found that effective use of classroom assessment yielded improvements in student achievement of between 0.4 and 0.7 standard deviations (although they noted problems with the interpretation of effect sizes across different studies—see discussion in next section).

In framing their review, Black and Wiliam (1998a) first presented a number of “examples in evidence that illustrated a number of features of effective formative assessment. Perhaps the most important feature they identified was that, to be effective, formative assessment had to be integrated into classroom practice, requiring a fundamental reorganization of classroom operations:

> 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. (Black & Wiliam, 1998a, p. 16)

Black and Wiliam (1998a) also noted that for assessment to function formatively, the feedback information had to be used, and thus the differential treatments that are incorporated in response to the feedback are at the heart of effective learning. Moreover, for these differentiated treatments to be selected appropriately, teachers need adequate models of how students will react to, and make use of, the feedback. As Perrenoud (1998) noted in his commentary on the Black and Wiliam paper, “…the feedback given to pupils in class is like so many bottles thrown into the sea. No one can be sure that the message they contain will one day find a receiver” (p. 87). The consequence of this is that the design of effective formative assessment cannot be detached from the learning milieu in which it is undertaken. The motivations and self-perceptions of students, and their assessment histories, will all be important influences on how feedback is received (Deci & Ryan, 1994).

In order to address the influences on how feedback is received, the Black and Wiliam (1998a) review examined the student perspective, the role of teachers, and some of the systems for the organization of teaching in which formative assessment was a major component. In drawing out implications for the policy and practice of formative assessment, they concluded:

> There does not emerge, from this present review, any one optimum model on which … policy might be based. What does emerge is a set of guiding principles, with the general caveat that the changes in classroom practice that are needed are central rather than marginal, and have to be incorporated by each teacher into his or her practice in his or her own way … That is to say, reform in this dimension will inevitably take a long time and need continuing support from both practitioners and researchers. (p. 62)

Most of the work reviewed above focused on school-age students up to the age of 18. Nyquist (2003) focused on studies of feedback in college-aged learners. He reviewed approximately 3000 studies of the effects of feedback, of which 86 met the following inclusion criteria:

> (a) experimental manipulation of a characteristic relevant to feedback;
> (b) used a sample of college-aged learners;
> (c) measured academic performance; and
> (d) provided sufficient quantitative information for an effect size to be calculated.

From the 86 studies it was possible to derive 185 effect sizes. The analysis yielded a mean effect size of 0.40 standard deviations—almost identical to that found by Kluger and DeNisi (1996). Weighting the effects so that their contribution to the mean effect was proportional to their reliability reduced this mean effect slightly to 0.35 (*SE* = 0.17), although the effects themselves were highly variable, ranging from -0.6 to 1.6 standard deviations. In order to investigate moderators of effect, Nyquist developed the following typology of different kinds of formative assessment:

> *Weaker feedback only*: students are given only the knowledge of their own score or grade, often described as "knowledge of results."
>
> *Feedback only*: students are given their own score or grade, together with either clear goals to work towards, or feedback on the correct answers to the questions they attempt, often described as "knowledge of correct results."
>
> *Weak formative assessment*: students are given information about the correct results, together with some explanation.
>
> *Moderate formative assessment*: students are given information about the correct results, some explanation, and some specific suggestions for improvement.
>
> *Strong formative assessment*: students are given information about the correct results, some explanation, and specific activities to undertake in order to improve.

Table 2 provides the average standardized effect size for each type of intervention, although these are corrected values that differ from those in the original thesis (J. B. Nyquist, personal communication, May 7, 2007). Nyquist's results echo the findings of Bangert-Drowns *et al*. (1991) discussed above. Just giving students feedback about current achievement produces only modest benefits, but where feedback engages students in mindful activity, the effects on learning can be profound. The effect sizes found by Nyquist also underscore the domain-specificity of effective feedback mentioned earlier.

Table 2: Effect Sizes for Different Kinds of Feedback Interventions (Nyquist, 2003)

	N	Effect Size
Weaker feedback only	31	0.14
Feedback only	48	0.36
Weaker formative assessment	49	0.26
Moderate formative assessment	41	0.39
Strong formative assessment	16	0.56
Total	185	

From the reviews of research conducted by Natriello (1987), Crooks (1988), Bangert-Drowns *et al*. (1991), and Black and Wiliam (1998a), it is clear that not all kinds of feedback to students about their work are equally effective. More recent research supports this assertion. For example, Meisels, Atkins-Burnett, Xue, Bickel and Son (2003) explored the impact of the Work Sample System (WSS)—a system of curriculum-embedded performance assessments—on the achievement of 96 third grade urban students in reading and mathematics, as measured by the Iowa Test of Basic Skills. When compared with a sample of 116 third graders in matched schools and with students in the remainder of the school district (Pittsburgh, PA), the achievement of WSS students was significantly and substantially higher in reading. In mathematics, however, the differences were much smaller, and failed to reach statistical significance. It would therefore appear that different school subjects may require different approaches, again reinforcing the domain-specificity of effective interventions.

Classroom assessment systems such as the Work Sampling System are also often designed primarily for summative purposes, to monitor and report on student progress, with their use to generate information for formative purposes often relegated to a subordinate role. In her review of the research literature on classroom assessment, Brookhart (2004) began by undertaking online searches with “classroom assessment” as a search term. Excluding hits related not relevant to K-12 education (for example, studies conducted in higher education settings) generated a total of 41 empirical studies with a focus on academic work in K-12 education. She concluded that classroom assessment occurs at the intersection of three teaching functions: instruction, classroom management, and assessment, and noted that the theory relevant to classroom assessment came from several different fields, including individual differences psychology, the study of groups, and educational measurement. She also noted that many of the studies she cited approached the phenomena under study from a single disciplinary perspective (often psychology) or were atheoretical inventories of classroom practice. Where studies had mixed two or more practical or theoretical perspectives, she concluded that “the resulting picture of classroom assessment was richer and more multidimensional” (p. 454).

While many of the studies included in the reviews discussed above focus on older students, it is apparent that students’ attitudes to learning are shaped by the feedback they receive from a very early age. In a year-long study of eight kindergarten and first grade classrooms in six schools in England, Tunstall and Gipps (1996a; 1996b) identified a range of roles played by feedback. Like Torrance and Pryor (1998), they found that much of the feedback given by teachers to students focused on socialization: “I’m only helping people who are sitting down with their hands up” (Tunstall & Gipps, 1996b p. 395). Beyond this socialization role, they identified four types of feedback on academic work (see Table 3). Type A included feedback that rewarded or punished the students for their work, such as students being allowed to leave for lunch early when they had done good work, or threatened with not being allowed to leave for lunch if they hadn’t completed assigned tasks. Type B feedback was also evaluative but, while type A feedback focused on rewards and sanction, type B feedback indicated the teacher’s level of approval, e.g. “I’m very pleased with you” or “I’m very disappointed in you today”.

In contrast to the evaluative feedback classified as types A and B, feedback classified as types C and D was *descriptive*. Type C focused on the adequacy of the work in terms of the teacher’s criteria for success, ranging from the extent to which the work already satisfied the criteria at one end (e.g., “This is extremely well explained”) to the steps the student needed to take to improve the work (e.g., “I want you to go over all of them and write your equals sign in each one”). A defining characteristic of type C feedback is that it focuses on the idea of work as product, while type D feedback emphasizes process aspects of work, with the teacher playing the role of facilitator, rather than evaluator. As Tunstall and Gipps (1996b) explain, teachers engaged in this kind of feedback “conveyed a sense of work in progress, heightening awareness of what was being undertaken and reflecting on it” (p. 399).

Table 3: Typology of Teacher Feedback (adapted from Tunstall & Gipps, 1996a)

Evaluative feedback	Type A	Type B
Positive	Rewarding	Approving
Negative	Punishing	Disapproving

Descriptive feedback	Type C	Type D
Achievement feedback	Specifying attainment	Constructing achievement
Improvement feedback	Specifying improvement	Constructing the way forward

From 2002 to 2004, as part of its research program on “What works in innovation in education” the Organisation for Economic Cooperation and Development (OECD) undertook a review of the practice of formative assessment in lower-secondary school classrooms in eight countries: Australia, Canada, Denmark, England, Finland, Italy, New Zealand and Scotland (Looney, 2005). As well as detailed case studies of the eight systems included in the review, the report of the project also contained reviews of the research on formative assessment published in French (Allal & Lopez, 2005) and German (Köller, 2005). Allal and Lopez reported that work by researchers in France and French-speaking parts of Belgium, Canada and Switzerland has focused much more on theoretical than empirical work, with very few controlled empirical studies. They suggest that the most important finding of the review of over 100 studies published in French over the last thirty years is that the studies of assessment practices in French speaking classrooms have utilized an “enlarged conception of formative assessment” (p. 245), along the lines adopted by Black and Wiliam (1998a).

In particular, Allal and Lopez argue that the central concept in the approach to feedback espoused within the Anglophone tradition, for example by Bloom, is that of “remediation,” which they summarize as “feedback+correction.” In contrast, within much of the research undertaken in Francophone countries, the central concept is “regulation”, summarized as “feedback+adaptation” (p. 245). It is important to note that the French word *régulation* has a much more specific meaning than the English word “regulation”. There are two ways to translate the word “regulation” into French—*règlement* and *régulation*. The former of these is used in the sense of “rules and regulations,” while the latter is used in the sense of adjustment in the way that a thermostat regulates the temperature of a room.

In their review, Allal and Lopez (2005) identify four major developments in the development of the conception of formative assessment in the French-language literature over the last thirty years. In the first, which they term “Focus on instrumentation” the emphasis was on the development of assessment tools such as banks of diagnostic items and adaptive testing systems. In the second, entitled “Search for theoretical frameworks”, the emphasis shifted to a “search for theories that can offer conceptual orientation for conducting assessment” (p. 249). The third development—“Studies of existing assessment practices in their contexts”—provides a grounding for the search for theoretical frameworks by articulating it with the study of how formative assessment is practiced in real classrooms. The fourth and most recent development has been “Development of active student involvement in assessment” which has examined student self-assessment, peer assessment, and the joint construction of assessment by students and teachers together.

The notion of formative assessment as being central to the regulation of learning processes has been adopted by some writers in the Anglophone community (see, for example, Wiliam, 2007), and the broadening of the conception of formative assessment in the English-language literature was noted by Brookhart (2007). Her review of the literature on "formative classroom assessment" charted the development of the conception of formative assessment as a series of nested formulations:

> *Formative assessment provides information about the learning process;*
>
> Formative assessment provides information about the learning process *that teachers can use for instructional decisions;*
>
> Formative assessment provides information about the learning process that teachers can use for instructional decisions *and students can use in improving their performance;*
>
> Formative assessment provides information about the learning process that teachers can use for instructional decisions and students can use in improving their performance, *which motivates students*.

In general, however, there would appear to be few links between the strong theoretical work in the Francophone tradition and the strong empirical work undertaken, particularly in the United States. Allal and Lopez (2005) concluded that "studies of practice are episodic and dispersed in different settings, which makes it difficult to identify patterns or trends. In summary, the theoretical promise of French-language work on formative assessment is in need of considerably more empirical grounding" (p. 256).

The review of the German-language literature by Köller (2005) began with an approach similar to that adopted by Black and Wiliam, with searches of on-line databases supplemented by scrutiny of all issues from 1980 to 2003 of the six most relevant German-language journals. Köller noted that, while there were many developments related to formative assessment reported in academic journals, there was little evaluation of the outcomes of formative assessment practices for students, although there were important confirmations of some findings in the Anglophone literature. Perhaps most notably, Köller reports the work of Meyer who, like Kluger and DeNisi, found that praise can sometimes have a negative impact on learning, while criticism, or even blame, can sometimes be helpful. Another important strand of work mentioned by Köller concerns differences between teachers in their use of reference norms. A number of studies, notably those by Rheinberg (1980), have shown that students learn more when taught by teachers who judge a student's performance against the same student's previous performance (an individual reference norm) rather than teachers who compare students with others in the class (a social reference norm).

Most recently, three substantial reviews on formative assessment have appeared. The first (Wiliam, 2007), focused specifically on mathematics education. As well as reviewing the research evidence on formative assessment, Wiliam drew out some of the implications of this research for mathematics teaching and outlined how the central ideas of formative assessment could be integrated within the broader idea of the regulation of learning processes developed from the French-language literature summarized above.

The other two recent reviews appeared in consecutive years in the journal *Review of Educational Research*. As part of a broader research program on the development of intelligent tutoring environments, Shute (2008) examined the research on feedback to students. A total of 141 publications met the inclusion criteria (103 journal articles, 24 books and book chapters, 10 conference proceedings and four research reports). While, as might be expected, Shute's review identified major gaps in the literature and concluded that there was no simple answer to the question, "What feedback works?", the review did endorse the findings of earlier reviews on the size of the effects that could be expected from feedback: standardized effect sizes ranged from 0.4 to 0.8 standard deviations. Shute also offered a number of preliminary guidelines for the design of effective feedback:

1. *Guidelines to enhance learning*. Feedback should focus on the specific features of the task, and provide suggestions on how to improve, rather than focus on the learner; it should focus on the "what, how and why" of a problem rather than simply indicating to students whether they were correct or not; elaborated feedback should be presented in manageable units and, echoing Einstein's famous dictum, should be "as simple as possible but no simpler." However, feedback should not be so detailed and specific that it scaffolds the learning so completely that the students do not need to think for themselves. Feedback is also more effective when from a trusted source (whether human or computer).

2. *Guidelines in relation to the timing of feedback*. The optimum timing of feedback appears to depend strongly on the kind of learning being undertaken. Immediate feedback appears to be most helpful for procedural learning, or where the task is well beyond the learner's capability at the beginning of the learning, while delayed feedback appears to be more appropriate for tasks well within the learner's capability, or where transfer to other contexts is sought.

A review by Hattie and Timperley (2007) summarizes an extensive program of work conducted by Hattie and his colleagues on systematic reviews of influences on student achievement. An earlier paper (Hattie, 1999) described the construction of a database of 500 meta-analyses, reporting 450,000 effect sizes from 180,000 studies involving over 20 million participants. An analysis of the 74 meta-analyses used in the 1999 study that specifically mentioned feedback found an average effect size of 0.56 across 13,370 effect sizes in the 74 meta-analyses (Hattie and Timperley, 2007), but Hattie and Timperley found, as had Kluger and DeNisi (1996), that there was significant variability amongst the various feedback studies in their effects on learning. The average of the 5755 effect sizes studies that Hattie and Timperley summarized as "Feedback" was 0.95 standard deviations, topped only by 89 studies coded as "Cues", which averaged 1.1 standard deviations.

Hattie and Timperley define the purpose of feedback as reducing discrepancies between current understandings or performance and a desired goal (as proposed by Ramaprasad, 1983). Building on the work of Deci and Ryan (1994) and Kluger and DeNisi (1996), their model posits that students can reduce the discrepancy either by employing more effective strategies or by increasing effort on the one hand, or by abandoning, blurring or lowering the goals they have set for themselves on the other hand. Teachers can reduce the discrepancy by changing the difficulty or the specificity of the goals, or by providing more support to the students. The model specifies three kinds of questions that feedback is designed to answer (Where am I going? How am I going? Where next?) and each feedback question operates at four levels: feedback about the task (FT), feedback about the processing of the task (FP), feedback about self-regulation (FR) and feedback about the self as a person (FS). They demonstrate that FS is the least effective form of feedback, that FR and FP "are powerful in terms of deep processing and mastery of tasks" (pp. 90-91) while FT is powerful when the feedback is used either to improve strategy processing, or for enhancing self-regulation (although they note that these conditions are rarely met in practice). The role of self-regulation in formative assessment is taken up in more detail below.

Definitions of Formative Assessment and Assessment for Learning

While the research reviewed above suggests that the use of assessment to inform instruction might have significant impact on learning, different reviews find very different effect sizes for the benefits of formative assessment. Kluger and DeNisi (1996) found an average effect size of 0.41 for feedback interventions, while Black and Wiliam (1998) estimated that the effects of formative assessment were around 0.4 to 0.7 standard deviations. Shute (2008) suggested a similar range (0.4 to 0.8) but Hattie and Timperley proposed an average effect size of 0.96 standard deviations for the effect of feedback. On the other hand, in a classroom setting, carried out over a year, with ordinary teachers, and where performance was measured using externally-mandated standardized tests, Wiliam, Lee, Harrison and Black (2004) found that a range of formative assessment strategies introduced by teachers had an effect size of 0.32 standard deviations. This is a substantial effect (the authors estimated this was equivalent to an increase of the rate of student learning of 70%, or an extra eight months of learning per year), but only one-third of the size of effects suggested by Hattie and Timperley.

Part of the variability is, no doubt, caused by differences in the sensitivity of the measures used in the different studies to the effects of instruction —see Wiliam (2010 pp. 20-22) for a discussion of the relationship between sensitivity to instruction and effect size. Effect sizes will also be affected by differences in the variance in the population. Many studies included in reviews of research are conducted on sub-populations that are not representative of the whole population. For example, if an effect size is calculated in a study of different interventions for students with special educational needs, then that effect size would not be generalizable to the whole population—where the population is more variable, the standard deviation that is used as the denominator in the calculation of the effect size is larger, leading to a smaller estimate of the effect size.

However, it seems likely that a significant part—perhaps even most—of the variability is caused by differences in how the ideas of formative assessment or assessment for learning were operationalized. As Bennett (2009) points out, in an important critical review of the field, one cannot be sure about the effects of such changes in practice unless one has an adequate definition of what the terms formative assessment and assessment for learning actually mean, and a close reading of the definitions that are provided suggests that there is no clear consensus about the meanings of the terms formative assessment and assessment for learning.

As noted above, Bloom appeared to conceptualize formative assessment as a combination of feedback and instructional correctives. Black and Wiliam (1998a) 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" (Black & Wiliam, 1998 p. 140)

A number of authors have proposed somewhat narrower definitions, most commonly by requiring the changes to instruction to take place during the instruction, as the following four quotations illustrate:

> "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, Hammerness, Darling-Hammond, Rust, Snowden, Gordon, Gutierrez & Pacheco, 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 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 summarized the characteristics of assessment that promotes learning:

> it is embedded in a view of teaching and learning of which it is an essential part;
>
> it involves sharing learning goals with pupils;
>
> it aims to help pupils to know and to recognise the standards they are aiming for;
>
> it involves pupils in self-assessment;
>
> it provides feedback which leads to pupils recognising their next steps and how to take them;
>
> it is underpinned by confidence that every student can improve;
>
> it involves both teacher and pupils reviewing and reflecting on assessment data. (Broadfoot, Daugherty, Gardner, Gipps, Harlen, James, & Stobart, 1999 p. 7)

In looking for a term to describe such assessments, they suggested that the term formative assessment was used in such different ways, that it 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, Daugherty, Gardner, Harlen, James, & Stobart, 2002 pp. 2-3).

The earliest use of the term *assessment for learning* appears to be a chapter of that title by Harry Black (1986). It was also the title of a paper given at AERA in 1992 (James, 1992)—the same year that a book called *Testing for learning* was published in the US (Mitchell, 1992)—and three years later, as the title of a book by Ruth Sutton (1995). In the United States, the origin of the term is often mistakenly attributed to Rick Stiggins as a result of his popularization of the term (see, for example, Stiggins, 2005), although Stiggins himself has always attributed the term to other authors.

Most recently, an international conference on assessment for learning in Dunedin in 2009, building on work done at two earlier conferences in the UK (2001) and the USA (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. (p. 264)

The phrase *assessment for learning* has an undoubted appeal, especially when contrasted with *assessment of learning*, but as Bennett (2009) points out, replacing one term with another serves merely to move the definitional burden. More importantly, as Black and Wiliam and their 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:

> 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)

Bennett (2009) endorses the idea that it is unhelpful, and simplistic, to equate assessment for learning with formative assessment and assessment of learning with summative assessment. From a "more nuanced" (p. 5) view, he suggests that assessments designed primarily to serve a summative function may also function formatively, while those designed primarily to serve a formative function may also function summatively. Consider the following seven assessment scenarios.

1. A team of mathematics teachers from the same school meet to discuss their professional development needs. They analyze the scores obtained by their students on national tests and see that while their scores are, overall, comparable to national benchmarks, their students tend to score less well on items involving ratio and proportion. They decide to make ratio and proportion the focus of their professional development activities for the coming year, meeting regularly to discuss the changes they have made in the way they teach this topic. Two years later, they find that their students are scoring well on items on ratio and proportion in the national tests, which takes their students' scores well above the national benchmarks.

2. Each year, a group of fourth-grade teachers meet together to review students' performance on a standardized reading test, and in particular, look at the facility (proportion correct) for different kinds of item on the test. Where item facilities are lower than expected, they look at how the instruction on that aspect of reading was planned and delivered, and they look at ways in which the instruction can be strengthened in the following year.

3. Every seven weeks, teachers in a school use a series of "interim" tests to check on student progress. Any student who scores below a threshold judged to be necessary to make adequate progress is invited to attend additional instruction. Any student who scores below the threshold on two successive occasions is *required* to attend additional instruction.

4. A teacher designs an instructional unit on *Pulleys and levers*. Following the pattern that is common in middle schools in Japan (Lewis, 2002 p. 76), although 14 periods are allocated to the unit, the teacher makes sure that all the content is covered in the first 11 periods. In period 12, the students complete a test on what they have covered in the previous 11 periods, and the teacher collects in the students responses, reads them, and, on the basis of what she learns about the class's understanding of the topic, plans what she is going to do in lessons 13 and 14.

5. A teacher has just been discussing with a class why historical documents cannot be taken at face value. As the lesson is drawing to a close, each student is given an index card (8cm by 13cm) and is asked to write an answer to the question "Why are historians concerned about bias in historical sources?" As they leave the classroom, the students hand the teacher these "exit passes" and after all the students have left, the teacher reads through the cards, and then decides how to begin the next lesson.

6. A sixth-grade class has been learning about different kinds of figurative language. In order to check on the class's understanding, the teacher gives each student a set of six cards bearing the letters A, B, C, D, and E. On the interactive white board, she displays the following list:

A. Alliteration
B. Onomatopoeia

C. Hyperbole
D. Personification
E. Simile

She then reads out a series of statements:

1. He was like a bull in a china shop.
2. This backpack weighs a ton.
3. He was as tall as a house
4. The sweetly smiling sunshine...
5. He honked his horn at the cyclist.

As each statement is read out to them, each member of the class has to hold up letter cards to indicate what kind of figurate language they have heard. The teacher realizes that almost all the students have assumed that each sentence can have only one kind of figurative language. She points out that the third sentence is a simile, but is also hyperbole, and she then re-polls the class on the last two statements, and finds that most students can now correctly identify the two kinds of figurative language in the last two statements. In addition, she makes a mental note of three students who answer most of the questions incorrectly, so that she can follow up with them individually at some later point.

7. A high-school chemistry teacher has been teaching a class how to balance chemical equations. In order to test the class, she writes up the unbalanced equation for the reaction of mercury hydroxide with phosphoric acid. She then invites students to change the quantities of the various elements in the equation, and when there are no more suggestions from the class, she asks the class to vote on whether the equation is now correct. All vote in the affirmative. The teacher concludes that the class has understood, and moves on.

In each of these seven scenarios, assessment information was used to make a better decision about instruction than would have been taken in the absence of the evidence. In the first two scenarios, the assessment instrument used had been designed entirely to serve a summative function, but the teachers involved found a way of using the evidence about student achievement elicited by the assessment to improve their instruction.

In the first six scenarios, the use of the evidence changed the instruction for the better while in the last, the assessment information confirmed that what the teacher had planned to do was indeed an appropriate course of action. In this sense, it was a better decision than it would have been in the absence of any evidence because it was better founded. In other words, evidence from assessments was used for the improvement of learning, even though many of the authors cited above would, in all probability not regard the first three as assessment for learning or formative assessment.

One response to this would be to try to restrict the meaning of formative assessment or assessment for learning to the kinds of assessment that are close to instruction, which would rule out the first three scenarios, and for some authors, the fourth also. However, restricting the meaning or use of the terms assessment for learning and formative assessment simply to try to ensure that the terms apply only to practices that are regarded favourably is rather perverse. It is rather like the approach used by a character in Lewis Carroll's *Through the Looking Glass*: "When *I* use a word … it means just what I choose it to mean — neither more nor less" (Carroll, 1871). The value of using the term *assessment* in phrases like *assessment for learning* and *formative assessment* is that it is illuminating to draw attention to the fact that the processes under consideration can be thought of assessment processes.

For this reason, it seems more helpful to acknowledge that in each of these cases, assessment was conducted with the intention of improving learning (although that may not have been the only reason for the assessment), and that the evidence from the assessments was used to improve instruction. For this reason, Black and Wiliam restated their original definition in a slightly different way, which they suggested was consistent with their original definition, and those others given above, including that of the Assessment Reform Group:

> 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)

A thorough exploration of the consequence of this definition is beyond the scope of this paper, but one point about this definition requires clarification. In explaining this definition, Black and Wiliam make clear that the term "instruction" is used in the sense in which it is used in the United States—the design of learning environments—and the "next steps in instruction" can be taken by teachers, learners, or their peers, or any combination of these three. The justification for each of the components of the definition can be found in Black and Wiliam (2009) and Wiliam (2010), and explorations of how formative assessment relates to other theoretical perspectives on teaching and learning can be found in Black and Wiliam (2004), Black and Wiliam (2011) and Wiliam (2007). In the final section of this paper, I explore some of the conditions that need to be in place for assessment to support learning.

When does assessment support learning?

While the definition proposed by Black and Wiliam above is relatively precise, it is much more a means for determining whether an assessment has, in fact, functioned formatively than it is a prescription for generating assessments that will, or are likely to, function formatively. From the research studied above, the two features that appear to be particularly important in designing assessment that will support learning is that the evidence generated is "instructionally tractable" (Wiliam, 2007). In other words, the evidence is more than information about the presence of a gap between current and desired performance. The evidence must also provide information about what kinds of instructional activities are likely to result in improving performance. For example, a low score on a mathematics test is likely to indicate nothing more than that the student has not yet learned what was intended. The only guidance this provides to the teacher is that more instruction is required. If the assessment has been designed to support valid inferences about specific aspects of performance, then the teacher might also realize that the student is having particular difficulties with rank ordering fractions. This allows the teacher to focus the remedial instruction more narrowly, but provides little insight into why the student is having difficulty. If, however, the assessment reveals a specific issue—for example that the student believes that only the size of the denominator matters when comparing fractions (Vinner, 1997) then this provides clear guidance for the teacher about what kinds of instructional activities to provide for the learner.

The second requirement is that the learner engages in actions to improve learning; this may be undertaking the remedial activities provided by the teacher, asking a peer for specific help, or reflecting on different ways to move her own learning forward—after all, the best designed feedback is useless if it is not acted upon. In other words, feedback cannot be evaluated without also taking into account the instructional context in which it is provided, and used. In the same way that engineers design feedback systems rather than simply ways of generating data, to understand feedback we must look at the learning milieu in which it operates. Thoughtful feedback given to students who have come to believe that they are "no good" at a particular subject is likely to be ignored or rejected, or appropriated in some other way to allow the learner to preserve a sense of well-being.

The involvement of learners, and their peers, was explicitly incorporated by Wiliam & Thompson (2008) in their proposal that formative assessment could be conceived of as involving three main processes (identifying where learners are in their learning, where they are going, how to get there) exercised by three categories of actors (teacher, learner, peer).

The resulting matrix of nine cells, they suggested, could be organized as five "key strategies" of formative assessment, as shown in figure 1.

	Where the learner is going	Where the learner is right now	How to get there
Teacher	Clarifying learning intentions and sharing and criteria for success	Engineering effective classroom discussions, activities and tasks that elicit evidence of learning	Providing feedback that moves learners forward
Peer	Understanding and sharing learning intentions and criteria for success	Activating learners as instructional resources for one another	
Learner	Understanding learning intentions and criteria for success	Activating learners as the owners of their own learning	

Figure 1: Aspects of Formative Assessment (Wiliam & Thompson, 2008)

While each of these five "key strategies" has generated a substantial research basis individually (see Wiliam, 2007, for a summary) they can also be viewed collectively as strategies for the regulation of learning processes. Indeed, Black and Wiliam (2009) point out that formative assessment involves "the creation of, and capitalization upon, 'moments of contingency' in instruction for the purpose of the regulation of learning processes" (p. 12).

Any attempt at the regulation of learning processes requires at least some idea of a goal, whether this is conceptualized as a single learning destination, or a broad "horizon" of learning goals any of which are equally acceptable. The teacher's role is then to elicit evidence of achievement, and undertake appropriate action to direct, or re-direct the learning in the intended direction. Within this formulation, the role of peers is analogous to that of teachers—while peers may lack the training and experience of teachers, they have unique insights into learning, and because the power relationships between peers are different from those between teachers and students, there will be instructional strategies open to them that would not be open, or would be less effective, when used by teachers.

The final strategy, “Activating students as owners of their own learning” clearly draws together a number of related fields of research, such as metacognition (Hacker, Dunlosky & Graesser, 1998), motivation (Deci & Ryan, 1994), attribution theory (Dweck, 2000), interest (Hidi & Harackiewicz, 2000) and, most importantly, self-regulated learning, defined by Boekaerts (2006) as “a multilevel, multicomponent process that targets affect, cognitions, & actions, as well as features of the environment for modulation in the service of one’s goals” (p. 347). While much of the research on self-regulation has tended to prioritize either cognitive or motivational approaches, in recent years there have been several significant attempts to draw these two strands more closely together, because, as Boekaerts (2006) argues, self-regulated learning is both metacognitively governed *and* affectively charged (p. 348).

Boekaerts has proposed a deceptively simple, but powerful, model for understanding self-regulated learning, termed the *dual processing* theory (Boekaerts, 1993). In the model:

> 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. (Boekaerts, 2006, p. 349)

Depending on the outcome of the appraisal, the student activates attention along one of two pathways: the “growth pathway” where the goal is to increase competence or the “well-being pathway” where attention is focused on preventing threat, harm or loss. While the former is obviously preferable, the latter is not necessarily counter-productive—by attending to the well-being pathway, the student may find a way to restore well-being (for example by lowering the cost of failure) that allows a shift of energy and attention to the growth pathway.

Students who are personally interested in a task are obviously likely to activate energy along the growth pathway, but where students are not personally interested in a task, a number of features of the task-in-context may nevertheless spark situational interest. Considerations of the trade-off between task *value* and *cost* will also influence how students direct their energies. In particular, students are more likely to focus on growth rather than well being when they see ability as incremental rather than fixed (Dweck, 2001), when they have a mastery rather than a performance orientation (Dweck, 2001) and when they identify with the goal (Deci & Ryan, 1994).

To summarize, because learning is unpredictable, assessment is necessary to make adaptive adjustments to instruction, but assessment processes themselves impact the learner’s willingness, desire, and capacity to learn (Harlen & Deakin-Crick, 2002). For assessment to support learning, it must provide guidance about the next steps in instruction and must be provided in way that encourages the learner to direct energy towards growth, rather than well-being.

Conclusion

The idea that assessment can support learning is not a new idea. It is inconceivable that those involved in the earliest attempts to communicate ideas, skills, or practices to others did not realize that such attempts could not be guaranteed to be successful, and that effective instruction therefore required evaluation, and adjustment. However, it is only forty years since Benjamin Bloom first suggested that it might be useful or illuminative to examine these processes as assessment. At the time, Bloom indicated that such processes would be more effective if they were separated from the use of assessment to record the achievement of learners, but for the next twenty years, the dominant role of assessment was seen as the recording of student achievement, although there were a number of attempts to use evidence collected for the purpose of summarizing achievement in other ways, notably for the improvement of instruction. However, it was not until the late 1980s that the idea that classroom assessment practices could both afford *and* constrain student learning began to gain widespread acceptance; used appropriately assessment could substantially improve learning, but that most of the time, the impact of assessment practices was to limit, and even to reduce, student learning.

During the 1990s, a number of studies explored the idea that attention to assessment as an integral part of instruction could improve learning outcomes for students, and at the same time, attempts were made to connect classroom practice to related bodies of research, notably feedback, motivation, attribution, and self-regulated learning. For most of this time, the term "formative assessment" was not precisely defined, and, as a result, research studies on one aspect of the use of assessment to improve instruction were used as evidence supporting the efficacy of quite unrelated aspects. Partly in response to this, many authors stopped using the term "formative assessment" preferring instead the phrase "assessment for learning" although again its precise meaning was rarely defined, beyond the idea that assessment should be used during instruction to improve learning outcomes.

This paper has reviewed these developments, and described more recent attempts that have been made to theorize formative assessment and assessment for learning in a number of ways, specifically in terms of classroom strategies and practical techniques that teachers can use to improve the quality of evidence on which the instructional decisions they, and their students, make. While there remains much more work to be done to integrate research on assessment for learning with more fundamental research on instructional design, feedback, self-regulated learning, and motivation, there is now a strong body of theoretical and empirical work that suggests that integrating assessment with instruction may well have unprecedented power to increase student engagement and to improve learning outcomes.

References

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, France: Organisation for Economic Cooperation and Development.

Ausubel, D. P. (1968). *Educational psychology: a cognitive view*. New York, NY: Holt, Rinehart & Winston.

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

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research,* **85**(2), 89-99.

Bennett, R. E. (2009). *A critical look at the meaning and basis of formative assessment* (ETS RM-09-06). Princeton, NJ: Educational Testing Service.

Black, H. (1986). Assessment for learning. In D. L. Nuttall (Ed.), *Assessing Educational Achievement.* (pp. 7-18). London: Falmer 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), 7-74.

Black, P. J., & Wiliam, D. (1998b). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan,* **80**(2), 139-148.

Black, P. J., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability,* **21**(1), 5-31.

Black, P., & Wiliam, D. (2011). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (2 ed.). London, UK: Sage.

Bloom, B. S. (1984a). The search for methods of instruction as effective as one-to-one tutoring. *Educational Leadership,* **41**(8), 4-17.

Bloom, B. S. (1984b). The 2-sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher,* **13**(6), 4-16.

Boekaerts, M. (1993). Being concerned with well being and with learning. *Educational Psychologist,* **28**(2), 149-167.

Boekaerts, M. (2006). Self-regulation and effort investment. In K. A. Renninger & I. E. Sigel (Eds.), *Handbook of child psychology volume 4: child psychology in practice* (6 ed., pp. 345-377). New York, NY: Wiley.

Broadfoot, P. M., Daugherty, R., Gardner, J., Gipps, C. V., Harlen, W., James, M., & Stobart, G. (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, NY: Teachers College Press.

Brophy, J. (1981). Teacher praise: a functional analysis. *Review of Educational Research,* **51**(1), 5-32.

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.

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

Elshout-Mohr, M. (1994). Feedback in self-instruction. *European Education,* **26**(2), 58-73.

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

Guskey, T. R. (2010). Formative assessment: the contributions of Benjamin S. Bloom. In H. L. Andrade & G. J. Cizek (Eds.), Handbook of formative assessment (pp. 106-124). New York, NY: Taylor & Francis.

Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.). (1998). Metacognition in educational theory *and practice*. Mahwah, NJ: Lawrence Erlbaum Associates.

Harlen, W., & Deakin-Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning. In EPPI-Centre (Ed.), *Research Evidence in Education Library* (1.1 ed., pp. 153). London, UK: University of London Institute of Education Social Science Research Unit.

Hattie, J. (1999, August 2). *Influences on student learning*. Retrieved August 25th, 2009, from http://www.education.auckland.ac.nz/uoa/education/staff/j.hattie/papers/influences.cfm

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

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*. Paper presented at the Annual Conference of the Association for Supervision and Curriculum Development (Assembly session on 'Critique of Reforms in Assessment and Testing in Britain') held at New Orleans, LA.

Kahl, S. (2005, 21 September). 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, France: Organisation for Economic Cooperation and Development.

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

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

Meisels, S. J., Atkins-Burnett, S., Xue, Y., Bickel, D. D., & Son, S.-H. (2003). Creating a system of accountability: the impact of instructional assessment on elementary children's achievement test scores. *Education Policy Analysis Archives,* **11**(9).

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

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 of Science thesis, Vanderbilt University.

Parkhurst, H. (1922). *Education on the Dalton Plan*. London, UK: G. Bell and Sons, Ltd.

Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning. Towards a wider conceptual field. *Assessment in Education: Principles Policy and Practice,* **5**(1), 85-102.

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

Reiser, R. A. (1986). Instructional technology: a history. In R. M. Gagné (Ed.), *Instructional technology: foundations* (pp. 11-48). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rheinberg, F. (1980). *Leistungsbewertung und lernmotivation [Achievement evaluation and learning motivation]*. Göttingen, Germany: Hogrefe.

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

Shepard, L. A., Hammerness, K., Darling-Hammond, L., Rust, F., Snowden, J. B., Gordon, E., Gutierrez, C., & Pacheco, A. (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.

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

Stiggins, R. J. (2005). From formative assessment to assessment FOR learning: a path to success in standards-based schools. *Phi Delta Kappan,* **87**(4), 324-328.

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment*. Buckingham, UK: Open University Press.

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

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

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 control and communication in the animal and the machine*. New York, NY: John Wiley & Sons Inc.

Wiliam, D. (2007). 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 Publishing.

Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 18-40). New York, NY: Taylor & Francis.

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. (2008). 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: Lawrence Erlbaum Associates.

Acknowledgement

Figure 1, which was first published in Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: minute-by-minute and day-by-day. *Educational Leadership,* **63**(3), 18-24, is used with permission of the Educational Testing Service.