Keeping learning on track: integrating assessment with instruction[1][2]

Dylan Wiliam, ETS[3]

Introduction

In this talk, I want to present some ideas about how assessment may play a fuller part not just in measuring the outcomes of learning, but in actually assisting the process of learning—in other words, a shift from assessment *of* learning to assessment *for* learning. The argument that I present rests on the following four assumptions.

- The primary purpose of educational research is the improvement of education. We cannot necessarily know which forms of research will pay dividends in the future so there is undoubtedly a place for 'pure' research in education. Nevertheless even the 'purest' form of research is not conducted in a vacuum, and while the implications of 'pure' research in education may not be immediately apparent, anything that helps us understand educational processes can help to illuminate or define the challenges that face us in improving education. In the language of Donald Stokes (1997) most educational research should be rooted firmly in 'Pasteur's quadrant'.

- The purpose of education is the improvement of student achievement. While we may argue about how this may be measured—and there is no doubt that some aspects of achievement are more easily measured than others—the purpose of education is to change learners; to enable them to do things that they could not previously do. In this sense, I have no problem with input-process-output models of education.

- The improvement of student achievement will be achieved primarily through changes in what happens in classrooms. Social changes such as improvements in diet, health and parenting will undoubtedly have their effects, but these will be slow compared to the changes that will be produced by changes in what happens in classrooms. Furthermore, while improvements in curricula, leadership and resources will also all help, they will help primarily by supporting more effective classroom practice. In this sense, an effective school is simply a school full of effective classrooms.

- The role of the teacher is not to teach *per se*, but rather to create situations in which students learn—in other words to 'engineer' learning environments. In those educational systems where there is pressure on teachers to raise the achievement of

---

[1] Invited address to the 30th annual conference of the International Association for Educational Assessment (IAEA) held in June 2004, Philadelphia, PA.

[2] I am grateful to Paul Black and Marnie Thompson for comments on an earlier draft of this paper.

[3] Address for correspondence: ETS, Rosedale Road (ms 04-R), Princeton, NJ 08541. Fax: (609) 734-1755; Email: dwiliam@ets.org

students, teachers frequently feel pressured to 'do more' to help their students learn. All too often, the result is that teachers take more and more of the responsibility for the students' learning, 'spoon-feeding' students the information that is needed to pass high-stakes assessments (Paris, Lawton, Turner & Roth, 1991). While it might appear obvious that this form of test preparation will help students pass tests, there is evidence that it is not the best way (Newmann, Bryk & Nagaoka, 2001, Nuthall & Alton-Lee, 1995), and that it may even be counter-productive (Boaler, 2002).

The logic of the foregoing argument is that the primary purpose of educational research is to assist in the creation of effective learning environments, and in this talk, I want to concentrate on the role that assessment can play in the design and operation of such effective learning environments. Specifically, I want to talk about the role that assessment plays in 'keeping learning on track' or more formally, in the regulation of learning.

In the next section, I discuss briefly the nature and purpose of assessments and review the research on the effects of classroom assessment. In the sections that follow, I outline the characteristics of effective classroom assessment and embed these ideas in the broader framework of the regulation of learning. Finally, I describe in more detail one study in which teachers began put these ideas into practice.

The nature and purpose of assessments

Educational assessments are conducted in a variety of ways and their outcomes can be used for a variety of purposes. There are differences in who decides what is to be assessed, who carries out the assessment, where the assessment takes place, how the resulting responses made by students are scored and interpreted, and what happens as a result. In particular, each of these can be the responsibility of those who teach the students, while at the other extreme, all can be carried out by an external agency. Cutting across these differences, there are also differences in the purposes that assessments serve. Broadly, education assessments serve three functions:

formative	supporting learning
summative	certifying individuals
evaluative	holding educational institutions to account

Through a series of historical contingencies, we have arrived at a situation in many countries in which the *circumstances* of the assessments have become conflated with the *purposes* of the assessment (Black and Wiliam, 2004a). So, for example, it is often widely assumed that the role of classroom assessment should be limited to supporting learning and all assessments with which we can hold educational institutions to account must be conducted by an external agency, even though in some countries, this is not the case.

In broad terms, moving from formative through summative to evaluative functions of assessment requires data at increasing levels of aggregation, from the individual to the institution, and from specifics of particular skills and weaknesses to generalities about overall levels of performance (although of course evaluative data may still be disaggregated in order to identify specific sub-groups in the population that are not making progress, or to identify particular weaknesses in students' performance in specific areas). However, it is also clear that the different functions that assessments may serve are in tension. The use of

data from assessments to hold schools accountable has, in many cases, because of 'teaching to the test', rendered the data almost useless for attesting to the qualities of individual students (apart, of course, from those qualities that are tested) or for supporting learning. Many authors have argued that these tensions require that one assessment system cannot serve all functions and that separate systems are required. No matter how convincing the argument in favor of this suggestion may be, it seems to me that it must not be believed because the consequences are so deleterious for learning. Separate assessment systems result either in the exclusion of teachers from summative assessments, or requiring them to operate parallel but distinct assessment systems for summative and formative functions, which almost always results in the marginalization of the formative function. If we are to develop integrated systems that can serve formative, summative and evaluative systems, the question that then arises is which functions should serve as the basis of the assessment. The position adopted in this paper is that the formative function should come first (see Black and Wiliam, 2004b, for a more detailed argument on this point). The main reason for this is that fine-scale data that have been collected to support learning can always be aggregated to provide information on students and on institutions, but aggregated summative data on students and institutions cannot generally be disaggregated to identify learning needs. Tensions in the different uses of the data will, of course, remain, but these can be ameliorated, even if they can't be entirely eradicated.

In a series of papers summarized in Newton (2003) and Wiliam (2003a) I have sketched out how an assessment system might be designed to serve all three functions reasonably well, but I also believe that formative assessment has a crucial role in *any* assessment regime. In other words, even if the teacher sees her or his task solely as the preparation of students for an external high-stakes test, formative assessment has a role to play.

Two substantial review articles, one by Gary Natriello (1987) and the other by Terry Crooks (1988) provided clear evidence that classroom evaluation practices have substantial impact on students and their learning. Natriello's review used a model of the assessment cycle, beginning with purposes, and moving on to the setting of tasks, criteria and standards, evaluating performance and providing feedback and then discusses the impact of these evaluation processes on students. His most significant point was that the vast majority of the research he cited was largely irrelevant because of weak theorization, which resulted in key distinctions (e.g. the quality and quantity of feedback) being conflated.

Crooks' paper had a narrower focus—the impact of evaluation practices on students. He concluded that the summative function of assessment has been too dominant and that more emphasis should be given to the potential of classroom assessments to assist learning.. Most importantly, assessments must emphasize the skills, knowledge and attitudes regarded as most important, not just those that are easy to assess.

However, the difficulty of reviewing relevant research in this area was highlighted by Black and Wiliam (1998a), in their synthesis of research published since the reviews by Natriello and Crooks. Those two papers had cited 91 and 241 references respectively, and yet only 9 references were common to both papers. In their own research, Black and Wiliam found that electronic searches based on keywords either generated far too many irrelevant sources, or omitted key papers, and in the end, they resorted to manual searches of each issue between 1987 and 1997 of 76 of the journals considered most likely to contain relevant research.

Black and Wiliam's review (which cited 250 studies) found that effective use of classroom assessment yielded improvements in student achievement between 0.4 and 0.7 standard deviations, and a recent review focusing on studies in higher education (Nyquist, 2003) found similar results.

There is therefore considerable evidence that attention to classroom assessment practices can have a substantial impact on student achievement. However, there is considerably less evidence about what are the key elements involved.

What is formative assessment?

In the United States, the term 'formative assessment' is often used to describe assessments that are used to provide information on the likely performance of students on state-mandated tests—a usage that might better be described as 'early-warning summative'. In other contexts it is used to describe any feedback given to students, no matter what use is made of it, such as telling students which items they got correct and incorrect (sometimes called 'knowledge of results'). These kinds of usages suggest that the distinction between 'formative' and 'summative' applies to the assessments themselves, but since the same assessment can be used both formatively and summatively, it follows that these terms cannot describe assessment themselves, but are really describing the use to which the resulting outcomes are put.

In some contexts, assessments that are used to support learning are described under the broad heading 'assessment for learning' (in contrast to 'assessment *of* learning'). This does suggest a process, rather than being a description of the nature of the assessment itself, but the danger here is that the focus is placed on the intention behind the use of the assessment, rather than action that actually takes place (Wiliam and Black, 1996). Many writers use the terms 'assessment for learning' and 'formative assessment' interchangeably, but Black et al (2002, p. i) distinguish between the two 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 pupils' 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 to be used as feedback, by teachers, and by their pupils, in assessing themselves and each other, to modify 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.

Another way of thinking about the distinction being made here is in terms of monitoring assessment, diagnostic assessment and formative assessment. An assessment *monitors* learning to the extent that it provide information about whether the student, class, school or system is learning or not; it is *diagnostic* to the extent that it provides information about what is going wrong; and it is *formative* to the extent that it provides information about what to do about it.

For the purpose of this talk, then, I take formative assessment to refer not to an assessment, nor even to the purpose of an assessment, but the function it actually serves. An assessment is formative to the extent that information from the assessment is fed back within the

system and actually used to improve the performance of the system in some way (i.e. that the assessment *forms* the direction of the improvement). For this to happen, Ramaprasad (1983) suggests that we need four things:

- information about the current state of the system
- information about the desired state of the system
- a way to determine whether there is a 'gap' between these two
- a mechanism whereby the feedback can be used to 'close the gap' between the current state and the goal state.

So, for example, if a student is told that she needs to work harder, and does work harder as a result, and consequently does indeed make improvements in her performance, this would *not* be formative. The feedback would be *causal*, in that it did trigger the improvement in performance, but not *formative*, because decisions about *how* to 'work harder' were left to the student. Telling students to 'Give more detail' might be formative, but only if the students knew what giving more detail meant (which is unlikely, because if they knew what detail was required, they would probably have provided it on the first occasion). Similarly, a 'formative assessment' that predicts which students are likely to fail the forthcoming state-mandated test is not formative unless the information from the test can be used to improve the quality of the learning within the system.

In order for assessment to function formatively, it needs to identify where learners are in their learning, where they are going, and how to get there. Crossing this three-fold typology of information needs with the different agents in the classroom (the student, her or his peers, and the teacher) creates the framework for looking at the role of formative assessment shown in figure 1. Figure 1 could be extended to include schools, districts or systems, but, as stated above, since the stance taken in this paper is that ultimately, assessment must feed into actions in the classroom in order to affect learning, this simplification seems reasonable.

To establish where the learner is in their learning, the teacher needs a range of ways of evoking information and eliciting the models that students hold (Lesh et al, 2003). This

	Teacher	Peer	Learner
Where the learner is	Evoking information	Peer-assessment	Self-assessment
Where she or he is going	Curriculum philosophy	Sharing success criteria	Sharing success criteria
How to get there	Feedback	Peer-tutoring	Self-directed learning

Figure 1: Aspects of formative assessment

can be through questions, other prompts (including statements to which the students have to react), or through problem situations which reveal the schemas with which the students are operating. For example, after showing students a spring balance suspending a weight inside a bell jar, they can be asked what will happen to the reading on the spring balance if the air inside the bell jar is evacuated, and to explain their reasoning. Many students respond that the weight will rise, because the air is no longer pressing down on it, suggesting that they believe that objects have weight only because of the air-pressure acting downwards on the

object (a belief no doubt reinforced by teachers' constant reminder of the 10 000 kg per square metre (or, in the US, 15 lbs per square inch) of air pressure pressing in on each of us, and of film showing astronauts weighing less on the airless moon than on earth).

Multiple-choice items also have a role to play here, provided they are constructed carefully. In traditional item-design, the purpose of the distractors is to 'distract' the weaker student from the correct answer, and provided that the resulting item has appropriate facility and discrimination, the item is regarded as satisfactory. The crucial point here is that in classical test theory, as well as most implementations of item-response theory, all incorrect responses are treated as equivalent in terms of information content. This is fine from a summative standpoint, but if we are to use items formatively, the distractors must be *interpretable*. In other words, from the observation of a student's choice of distractor, we need to be able to make inferences about the schemas that the student has used in arriving at that choice.

For example, if students are asked to provide the general term for the sequence

3, 7, 11, 15, …

we could provide the following choices:

(A) $3 + n$
(B) $n + 4$
(C) $3n + 4$
(D) $4n – 1$

The important feature of this item is that that the distractors are not generated randomly, but relate to well known misconceptions that students have about algebra. The first tends to be chosen by students who believe that the 'rule' is "start with 3 and keep on adding the same number" while the second is chosen by students with a similar misconception where the rule is "add 4 to the last number". Both of these misconceptions are associated with the idea, repeated often by teachers, but frequently misunderstood by students, that "n can be any number". The third choice identifies the "letter ignored" strategy identified by the Concepts in Secondary Mathematics and Science (CSMS) project (Hart, 1981).

This item includes just one correct response, and for high-stakes items, this is an important requirement, but for formative purposes, the item could be modified to include a second, correct, but less 'mathematical', choice:

(E) $4n + 3$

While correct, in that the expression does generate the sequence, it is less 'mathematical' than option (D) because the first term corresponds to $n = 0$ and the second to $n = 1$ and so on, while (D) has the first term as $n = 1$, the second as $n = 2$ and so on. Option (E) would widely be perceived as unfair if included alongside option (D) in an item in a high-stakes test, but would be a very useful way of generating classroom discussion. While some items might serve both summative and formative functions well, others may be suitable for only one of the two functions

By themselves, however, even good items are not enough. In order to function formatively, items must not only diagnose areas of weakness; they must also connect to instructional steps that can be taken to overcome the weakness. No matter how precisely an item identifies student misconceptions or weakness, this will be of little use unless teachers can interpret the responses made by students in terms of *learning needs*. In other words, responses, when interpreted appropriately, must also provide guidance for effective action.

In all this, it is important to note that not all model-eliciting activities are equally important, even if they do tell us something new about students' conceptions. The choice of which models to elicit must be driven by a clear philosophy of the subject. For a given curriculum, some things are important to know and some are not and so it is also necessary to be clear about the desired outcome of the learning. In some cases, this may be a specific goal (e.g. getting the students to be able to find the area of a trapezoid, or balance a chemical equation) but in the case of many aspects of the language arts and social studies, as well as in open-ended and exploratory work in mathematics and science, there may be a whole range of goals that are appropriate for different learners or for learners at different stages of development.

Such prompts can, as well as telling us where students are in their learning, also actually produce learning (assessment *as* learning). For example, students who have become familiar with the notion of heat energy might be asked to estimate the heating requirements of a swimming pool of given volume. Such a task guides students towards the invention for themselves of the notion of specific heat capacity of water (i.e. the amount of energy needed to heat a kilogram of water by one degree). Such 'big questions' (*hatsumon*) are a very powerful feature of lesson design in Japanese classrooms.

Once the learning outcomes are clear, the provision of feedback from the teacher can assist learning, provided, of course, such feedback is acted upon. Several conditions need to be met for this to take place. The feedback itself needs to be task-involving rather than ego-involving (Kluger & DeNisi, 1996), but it also helps if students see the purpose of feedback as helping them improve, rather than simply judging their worth, if the students have mastery, rather than performance goals, and see ability as incremental rather than fixed (Dweck, 2000). Too often. as Perrenoud (1998) notes, "…feedback given to pupils in class is like so many bottles thrown out to the sea. No one can be sure that the message they contain will one day find a receiver" (p87).

Learning is also enhanced when learners are able to assess their own performance (Fontana & Fernandes, 1994). But as Sadler (1989) notes, this requires that learners come to understand the criteria for success that the teacher already has in mind. Learners often find this difficult, however, and the involvement of peers can help learners understand success criteria and monitor their own progress towards their goals (White & Frederiksen, 1998). Thus peer-assessment provides an important complement to, and may even be a pre-requisite for, effective self-assessment (Black, Harrison, Lee, Marshall, & Wiliam, 2003).

Although the starting point for work on formative assessment was the relatively simple idea of feedback, the formulation above presents rather a complex picture of formative assesssment, and the ways in which the elements within figure 1 relate to each other is not straightforward. However, all the elements in figure 1 can be integrated within a more general theoretical framework of the *regulation of learning processes* as suggested

Perrenoud (1991, 1998)[4]. Within such a framework, the actions of the teacher, the learners, and the context of the classroom are all evaluated with respect to the extent to which they contribute to guiding the learning towards the intended goal.

Formative assessment and the regulation of learning

The first thing to say here is that it is important to distinguish between the regulation of the activity in which the student engages and the regulation of the learning that results. Most teachers appear to be quite skilled at the former, but have only a hazy idea of the learning that results. This is especially evident in interviews before lessons where teachers focus much more on the planned activities than on the resulting learning (e.g. "I'm going to have them do X"). In a way, this is inevitable, since only the activities can be manipulated directly. Nevertheless, it is clear that in teachers who have developed their formative assessment practices, there is a strong shift in emphasis from regulating the activities and towards the learning that results (Black et al, 2003). Indeed, from such a perspective, even to describe the task of the teacher as teaching is misleading, since it is rather to 'engineer' situations in which student learn.

The second point to make is that the 'engineering of learning environments' does not guarantee that the learning is well-regulated. Many visual arts classroom are *productive*, in that they do lead to significant learning on the part of students, but what any given student might learn is impossible to predict.

When the learning environment is well-regulated, much of the regulation is achieved 'upstream' of the lesson itself (i.e. before the lesson begins), through the setting up of didactical situations. The regulation can be unmediated within such didactical situations, when, for example, a teacher "does not intervene in person, but puts in place a 'metacognitive culture', mutual forms of teaching and the organisation of regulation of learning processes run by technologies or incorporated into classroom organisation and management" (Perrenoud, 1998 p100). For example, a teacher's decision to use realistic contexts in the mathematics classroom can provide a source of upstream regulation, because then students can determine the reasonableness of their answers. If students calculate that the average cost per slice of pizza (say) is $200, provided they are genuinely engaged in the activity, they will know that this solution is unreasonable, and so the use of realistic settings provides a 'self-checking' mechanism.

On the other hand, the didactical situation may be set up so that the regulation is achieved through the mediation of the teacher, when the teacher, in planning the lesson, creates questions, prompts or activities that evoke responses from the students that the teacher can use to determine the progress of the learning, and if necessary, to make adjustments. Examples of such questions are, "Is calculus exact or approximate?", "What is the pH of 10 molar NaOH?", or, "Would your mass be the same on the moon?". (In this context it is

---

[4] In English, the noun 'regulation' has two meanings; one refers to the act of regulating and the other to a rule or law to govern conduct, and so, while it is the former sense that is intended here, the word has the unfortunate connotation of the second. In French, the two senses have separate terms (régulation and règlement) and so the problem does not arise.

worth noting that each of these questions is 'closed' in that there is only one correct response—their value is that although they are closed, each question is focused on a specific misconception.)

The 'upstream' planning therefore creates, downstream, the possibility that the learning activities may change course in the light of the students' responses. These 'moments of contingency'—points in the instructional sequence when the instruction can proceed in different directions according to the responses of the student—are at the heart of the regulation of learning.

These moments arise continuously in whole-class teaching, where teachers are constantly having to make sense of students' responses, interpreting them in terms of learning needs, and making appropriate responses. But they also arise when the teacher circulates around the classroom, looking at individual students' work, observing the extent to which the students are 'on track'. In most teaching of mathematics and science, the regulation of learning will be relatively tight, so that the teacher will attempt to 'bring into line' all learners who are not heading towards the particular goal sought by the teacher—in these subjects, the *telos* of learning is generally both highly specific and general to all the students in a class. In contrast, in much teaching in language arts and social studies, the regulation will be much looser. Rather than a single goal, there is likely to be a broad *horizon* of appropriate goals, all of which are acceptable, and the teacher will intervene to bring the learners 'into line' only when the trajectory of the learner is radically different from that intended by the teacher. In this context, it is worth noting that there are significant cultural differences in how to use this information. In the United States, the teacher will typically intervene with individual students where they appear not to be 'on track' whereas in Japan, the teacher is far more likely to observe all the students carefully, while walking round the class, and then will select some major issues for discussion with the whole class.

One of the features that makes a lesson 'formative', then, is that the lesson can change course in the light of evidence about the progress of learning. This is in stark contrast to the 'traditional' pattern of classroom interaction, exemplified by the following extract:

> "Yesterday we talked about triangles, and we had a special name for triangles with three sides the same. Anyone remember what it was? … Begins with E … equi-…"

In terms of formative assessment, there are two salient points about such an exchange. First, little is contingent on the responses of the students, except how long it takes to get on to the next part of the teacher's 'script', so there is little scope for 'downstream' regulation. The teacher is interested only in getting to the word 'equilateral' in order that she can move on, and so all incorrect answers are treated as equivalent. The only information that the teacher extracts from the students' responses is whether they can recall the word 'equilateral' or not. This echoes the points made about classical test theory made above. In classical test theory, all incorrect responses are regarded as equivalent in terms of information content, and much teacher questioning treats all incorrect responses as equivalent in terms of information content; all the teacher learns is that the students didn't 'get' it.

The second point is that the situation that the teacher set up in the first place—the question she chose to ask—has little potential for providing the teacher with useful information

about the students' thinking, except, possibly, whether the students can recall the word 'equilateral'. This is typical in situations where the questions that the teacher uses in whole-class interaction have not been prepared in advance (in other words, when there is little or no 'upstream' regulation).

Similar considerations apply when the teacher collects in the students' notebooks and attempts to give helpful feedback to the students in the form of comments on how to improve rather than grades or percentage scores. If sufficient attention has not been given 'upstream' to the design of the tasks given to the students, then the teacher may find that she has nothing useful to say to the students. Ideally, from examining the students' responses to the task, the teacher would be able to judge how to (a) help the learners learn better and (b) what she might do to improve the teaching of this topic. In this way, the assessment could be formative for the students, through the feedback she provides, and formative for the teacher herself, in that appropriate analysis of the students responses might suggest how the lesson could be improved.

Assessments can also be formative at the level of the school, district, and state provided the assessments help to regulate learning. Frequent assessment can identify students who are not making as much progress as expected (whether this expectation is based on some notion of 'ability', prior achievement, or external demands made by the state). But frequent summative testing—we might call this micro-summative—is not formative unless the information that the tests yield is used in some way to modify instruction (see next section).

System responsiveness and time-frames

A key issue in the design of assessment systems, if they are to function formatively as well as summatively, is the extent to which the system can respond in a timely manner to the information made available. Feedback loops need to be designed taking account of the responsiveness of the system to the actions that can be used to improve its performance. The less responsive the system, the longer the feedback loops need to be for the system to be able to react appropriately.

For example, analysis of the patterns of student responses on a 'trial run' of a state-mandated test in a given school district might indicate that the responses made by students in seventh grade on items involving (say) probability were lower than would be expected given the students' scores on the other items, and lower than the scores of comparable students in other districts. One response to this could be a program of professional development on teaching probability for the seventh grade mathematics teachers in that district. Since this would take some weeks to arrange, and even longer for it to have an effect, the 'trial run' would need to be held some months before the state-mandated test in order to provide time for the system to interpret the data in terms of the system's needs. The 'trial run' would be formative for the district if, and only if, the information generated were used to improve the performance of the system—and if the data from the assessment actually helped to form the direction of the action taken.

For an individual teacher, the feedback loops can be considerably shorter. A teacher might look through the same students' responses to a 'trial run' of a state test and re-plan the topics that she is going to teach in the time remaining until the test. Such a test would be useful as little as a week or two before the state-mandated test, as long as there is time to

use the information to re-direct the teaching. Again this assessment would be formative as long as the information from the test was actually used to adapt the teaching, and in particular, not only telling the teacher which topics need to be re-taught, but also to suggest what kinds of re-teaching might produce better results.

The building-in of time for responses is a central feature of much elementary and middle school teaching in Japan. A teaching unit is typically allocated 14 lessons, but the content usually occupies only 10 or 11 of the lessons, allowing time for a short test to be given, and for the teacher to re-teach aspects of the unit that were not well-understood.

Another example, on an even shorter time-scale, is the use of 'exit passes' from a lesson. The idea here is that before leaving a classroom, each student must compose an answer to a key question given by the teacher at the end of the lesson. On a lesson on probability for example, such a question might be, "Why can't a probability be greater than one?" Once the students have left, the teacher can look at the students' responses, and make appropriate adjustments in the plan for the next period of instruction.

The shortest feedback loops are those involved in the day-to-day classroom practices of teachers, where teachers adjust their teaching in the light of students' responses to questions or other prompts in 'real time'. The key point in all this that the length of the feedback loop should be tailored according to the ability of the system to react to the feedback.

However, this does not mean that the responsiveness of the system cannot be changed. Through appropriate 'upstream' regulation, the responsiveness can be enhanced considerably. Where teachers have collaborated to anticipate the responses that students might make to a question, and what misconceptions would lead to particular incorrect responses, for example through the process of Lesson Study practiced in Japan (Lewis, 2002), the teachers would be able to adapt their instruction much more quickly, even to the extent of having alternative instructional episodes ready. In this way, feedback to the teacher that, in the normal course of things, might need at least a day to be used to modify instruction, could affect instruction immediately.

In the same way, a school district or state that has thought about how it might use the information about student performance before the students' results are available (for example by the preparation of particular kinds of diagnostic reports—see Wiliam, 1999) is likely to reduce considerably the time needed to use the information to improve instruction. As in other examples, attention to regulation 'upstream' pays dividends 'downstream'.

Putting it into practice

No matter how elegantly we formulate our ideas about formative assessment, they will be moot unless we can find ways of supporting teachers in incorporating more attention to assessment in their own practice. There are, of course, other ways that educational research can influence practice, such as through the design of curricula and textbooks, although as Clements (2002) notes, these impacts are generally small. If educational research is to have any lasting impact on practice, it must be taken up and used by practitioners.

Traditionally, researchers have engaged in a process of 'disseminating' their work to teachers, or engaging in 'knowledge transfer'. Both of these metaphors have some utility,

but they suggest that all researchers need to do is to "share the results" (English, Jones, Lesh, Tirosh, & Bussi, 2002, p. 805) of their research with practitioners and the findings will somehow be used.

However, the emerging research on expertise shows that the process of 'knowledge transfer' cannot be one of providing instructions to novices in the hope that they will get better (see Wiliam, 2003 for more on this point), because, put simply, all research findings are generalizations and as such are either too general to be useful, or too specific to be universally applicable. For example, the research on feedback such as the work of Kluger & DeNisi cited above suggests that task-involving feedback is to be preferred to ego-involving feedback, but what the teacher needs to know is, "Can I say, "Well done" to this student, now?" Put crudely, such generalizations underdetermine action.

At the other extreme, experts can often see that a particular recipe is inappropriate in some circumstances, although because their reaction is intuitive, they may not be able to discern the reason why. The message received by the practitioner in such cases is that the findings of educational research are not a valid guide to action.

The difficulty of 'putting research into practice' is the fault neither of the teacher nor of the researcher. Because our understanding of the theoretical principles underlying successful classroom action is weak, research cannot tell teachers what to do. Indeed, given the complexity of classrooms, it seems likely that the positivist dream of an effective theory of teacher action—which would spell out the 'best' course of action given certain conditions—is not just difficult and a long way off, but impossible in principle (Wiliam, 2003b).

What is needed instead is an acknowledgement that what teachers do in 'taking on' research is not a more or less passive adoption of some good ideas from someone else but an active process of knowledge *creation*:

> Teachers will not take up attractive sounding ideas, albeit based on extensive research, if these are presented as general principles which leave entirely to them the task of translating them into everyday practice—their classroom lives are too busy and too fragile for this to be possible for all but an outstanding few. What they need is a variety of living examples of implementation, by teachers with whom they can identify and from whom they can both derive conviction and confidence that they can do better, and see concrete examples of what doing better means in practice. (Black & Wiliam, 1998, p. 15)

For these reasons when Paul Black and I began work with teachers on formative assessment, we decided that we had to work in a genuinely collaborative way with a small group of teachers, suggesting directions that might be fruitful to explore, and supporting them as well as we could, but avoiding the trap of dispensing 'tips for teachers'. At first, it seems likely that the teachers did not believe this. They seemed to believe that we were operating with a perverted model of discovery learning in which we knew full well what we wanted the teachers to do, but didn't tell them, because we wanted the teachers 'to discover it for themselves'. However, after a while, it became clear that there was no prescribed model of effective classroom action, and each teacher would need to find their own way of implementing these general principles in their own classrooms.

Our model for working with teachers is based on what we know about what constitutes effective teacher professional development and in particular that it needs to attend to both

process and *content* elements (Reeves, McCall, & MacGilchrist, 2001; Wilson & Berne, 1999). On the process side, professional development is more effective when it is related to the local circumstances in which the teachers operate (Cobb, McClain, Lamberg, & Dean, 2003), takes place over a period of time rather than being in the form of one-day workshops (Cohen & Hill, 1998), and involves teacher in active, collective participation (Garet, Birman, Porter, Desimone, & Herman, 1999). In addition to these process elements, however, professional development is more effective when it has a focus on deepening teachers' knowledge of the content they are to teach, the possible responses of students, and strategies that can be utilized to build on these (Supovitz, 2001).

The details of how we worked with the teachers can be found in Black, Harrison, Lee, Marshall and Wiliam (2003), but is summarized here for ease of reference. The intervention had two main components:

> a series of six one-day inservice sessions, during which teachers would be introduced to our view of the principles underlying formative assessment, and have a chance to develop their own plans;
>
> visits to the schools, during which the teachers would be observed teaching by project staff, have an opportunity to discuss their ideas, and plan how they could be put into practice more effectively.

The key feature of the inset sessions was the development of action plans. Since we were aware from other studies that effective implementation of formative assessment requires teachers to re-negotiate the 'learning contract' (cf Brousseau, 1984) that they had evolved with their students, we decided that implementing formative assessment would best be done at the beginning of a new school year. For the first six months of the project, therefore, we encouraged the teachers to experiment with some of the strategies and techniques suggested by the research, such as rich questioning, comment-only marking (grading), sharing criteria with learners, and student peer- and self-assessment. Each teacher was then asked to draw up, and later to refine, an action plan specifying which aspects of formative assessment they wished to develop in their practice and to identify a focal class with whom these strategies would be introduced in the following September.

Most of the teachers' plans contained reference to two or three important areas in their teaching where they were seeking to increase their use of formative assessment, generally followed by details of strategies that would be used to make this happen. In almost all cases the plan was given in some detail, although many teachers used phrases whose meanings differed from teacher to teacher (even within the same school).

Practically every plan contained some reference to focusing on or improving the teacher's own questioning techniques although only about half gave details on how they were going to do this (for example using more open questions, allowing students more time to think of answers or starting the lesson with a focal question). Others were less precise (for example using more sustained questioning of individuals, or improving questioning techniques in general). Some teachers mentioned planning and recording their questions. Many teachers also mentioned involving students more in setting questions (for homework, or for each other in class)**.** Some teachers also saw existing national curriculum tests as a source of good questions.

Using comment-only marking was specifically mentioned by nearly half the teachers, although only 6 of the teachers included it as a specific element in their action plans. Some of the teachers wanted to reduce the use of scores and grades, but foresaw problems with this, given school policies on grading. Four teachers planned for a module test to be taken before the end of the module thus providing time for remediation.

Sharing the objectives of lessons or topics was mentioned by most of the teachers, through a variety of techniques (using a question that the students should be able to answer at the end of the lesson, stating the objectives clearly at the start of the lesson, getting the students to round up the lesson with an account of what they had learned). About half the plans included references to helping the students understand the grading criteria (rubrics) used for investigative or exploratory work, generally using exemplars from students from previous years. Exemplar material was mentioned in other contexts such as having work on display and asking students to mark work using a set of criteria provided by the teacher.

Almost all the teachers mentioned some form of self-assessment in their plans, ranging from using red, amber or green 'traffic lights' to indicate the student's perception of the extent to which a topic or lesson had been understood, to strategies that encouraged self-assessment via targets which placed responsibility on students (eg "One of these twenty answers is wrong: find it and fix it!"). Traffic lights (or smiley faces—an equivalent that did not require colored pens or pencils!) were seen in about half of the plans and in practically all cases their use was combined with strategies to follow up the cases where the students signaled incomplete understanding.

When we attempted to see whether particular combinations of strategies were selected by teachers, we could find no discernible patterns. Each teacher's choice of techniques to develop appeared to be entirely idiosyncratic.

The other component of the intervention, the visits to the schools, provided an opportunity for project staff to discuss with the teachers what they were doing, and how this related to their efforts to put their action plans into practice. The interactions were not directive, but more like a holding up of a mirror to the teachers. Since project staff were frequently seen as 'experts' in either mathematics or science education, there was a tendency sometimes for teachers to invest questions from a member of the project team with a particular significance, and for this reason, these discussions were often more effective when science teachers were observed by mathematics specialists, and vice-versa.

A detailed description of the qualitative changes in teachers' practices can be found in Black, Harrison, Lee, Marshall & Wiliam (2003), but it is worth noting here that the teachers' practices were slow to change, and that most of the changes in practice that we observed occurred towards the end of the year, so that the actual size of the effects found are likely to be underestimates of what could be achieved when teachers are emphasizing formative assessment as an integral part of their practice.

Since each teacher was free to decide the teaching group on which their development efforts would be focused, it was not possible to impose a standard experimental design. Instead, we identified, with the help of each teacher, the best possible comparison group, and set up a 'mini-experiment' for each teacher. The comparison of the achievements of the students in the focal group with the local comparison group indicates that the students

taught by teachers developing their use of formative assessment out-performed the comparison groups by approximately 0.3 standard deviations, as measured by external tests and examinations (see Wiliam, Lee, Harrison & Black, 2004 for details).

This work suggests that the effects found in the studies reviewed by Natriello, Crooks, and Black and Wiliam, most of which were conducted over a relatively short time-scale, are sustainable over the long-term. Almost all the teachers involved in the original project have continued to develop their formative assessment practices, and many are now outstanding practitioners. Perhaps more importantly, these teachers have also said that they enjoy their teaching more than they did previously, and have managed to enthuse other teachers about the potential of formative assessment to improve student learning.

The major problem in all this is that, while we know that the steps the teachers took were effective in improving student achievement, we don't know exactly what they did. Because we allowed each teacher to take on these ideas in his or her own way, and because each teacher modified their initial plan over the course of the year (often after hearing of the experiences of other teachers) we cannot determine whether some of the strategies used were more powerful than others. We know that the intervention was successful, but we don't know what it was! Much therefore remains to be done, but this experiences suggests that the development of formative assessment is a powerful lever for teacher professional devlopment, and the improvement of student achievement.

Conclusion

In this talk, I have argued that the terms formative and summative apply not to assessments themselves, but to the functions they serve, and as a result, the same assessment can be both formative and summative. Assessment is formative when the information arising from the assessment is fed back within the system and is actually used to improve the performance of the system. Assessment is formative for individuals when they can use the feedback from the assessment to improve their learning. Assessment is formative for teachers when the outcomes from the assessment, appropriately interpreted, help them improve their teaching, either on specific topics, or generally. Assessments are formative for schools and districts if the information generated can be interpreted in such a way as to improve the quality of learning within the schools and districts. The view of assessment presented here involves a shift from quality control in learning to quality assurance in learning. Rather than teaching students, and then, at the end of the teaching, finding out what has been learned, it seems obvious that what we should do is to assess the progress of learning whilst it is happening, so that we can adjust the teaching if things are not working. In order to achieve this, the length of the cycle from evidence to action must be designed taking into account the responsiveness of the system. Some feedback loops, such as those in the classroom, will be only fractions of a second long, while others, such as those involving districts or state systems will last months, or even years.

More generally, I have suggested, building on the work of Philippe Perrenoud, that formative assessment be considered as a key component of well-regulated learning environments. From this perspective, the task of the teacher is to not necessarily to teach, but rather to engineer situations in which students learn effectively. One way to do this is to design the environment so that the regulation is embedded within features of the environment. Alternatively, when the regulation is undertaken through the teacher's

mediation, it is necessary to build opportunities for such mediation into the instructional sequence by designing in episodes that will elicit students' thinking (upstream regulation) and to use the evidence from these probes to modify the instruction (downstream regulation).

Work with teachers suggests that the development of teachers' formative assessment practices is manageable and relatively inexpensive to implement. However, the changes are slow to take effect, and it is not yet clear how the model used here could be scaled up effectively. We have begun to explore what, exactly, changes when teachers develop formative assessment (Black and Wiliam, under review) , but much more remains to be done. In particular, we do not know whether some of these strategies have greater leverage than others, both for promoting professional development and increasing student achievement. Nevertheless, there are clearly reasons to be optimistic. Perhaps one day we will not talk about "integrating instruction with assessment" because the distinction between the two will be meaningless.

References

Black, P. J. & Wiliam, D. (2004a). Classroom assessment is not (necessarily) formative assessment (and vice-versa). In M. Wilson (Ed.) *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education (part 2)* Chicago, IL: University of Chicago Press.

Black, P. J. & Wiliam, D. (2004b). The formative purpose: assessment must first promote learning. In M. Wilson (Ed.) *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education (part 2)* Chicago, IL: University of Chicago Press.

Black, P. J. & Wiliam, D. (under review). Theory and practice in the development of formative assessment. Submitted for publication in *Educational Assessment*.

Black, P.; Harrison, C.; Lee, C.; Marshall, B. & Wiliam, D. (2002). *Working inside the black box: assessment for learning in the classroom*. London, UK: King's College London Department of Education and Professional Studies.

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

Boaler, J. (2002). *Experiencing school mathematics: traditional and reform approaches to teaching and their impact on student Learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Brousseau, G. (1984). The crucial role of the didactical contract in the analysis and construction of situations in teaching and learning mathematics. In H.-G. Steiner (Ed.) *Theory of mathematics education: ICME 5 topic area and miniconference* (pp. 110-119). Bielefeld, Germany: Institut für Didaktik der Mathematik der Universität Bielefeld.

Clements, D. H. (2002). Linking research and curriculum development. In L. D. English (Ed.) *Handbook of international research in mathematics education* (pp. 599-630). Mahwah, NJ: Lawrence Erlbaum Associates.

Cobb, P.; McClain, K.; Lamberg, T. d. S. & Dean, C. (2003). Situating teachers' instructional practices in the institutional setting of the school and district. *Educational Researcher,* **32**(6), 13-24.

Cohen, D. K. & Hill, H. C. (1998). *State policy and classroom performance: mathematics reform in California*. Philadelphia, PA: University of Pennsylvania Consortium for Policy Research in Education.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research,* **58**(4), 438-481.

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

English, L. D.; Jones, G.; Lesh, R.; Tirosh, D. & Bussi, M. B. (2002). Future issues and directions in international mathematics education research. In L. D. English (Ed.) *Handbook of international research in mathematics education* (pp. 787-812). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Garet, M. S.; Birman, B. F.; Porter, A. C.; Desimone, L. & Herman, R. (1999). *Designing effective professional development: lessons from the Eisenhower Program*. Washington, DC: US Department of Education.

Hart, K. M. (Ed.) (1981). *Children's understanding of mathematics: 11-16*. London, UK: John Murray.

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.

Lesh, R.; Hoover, M.; Hole, B.; Kelly, A. E. & Post, T. (2003). Principles for developing thought revealing activities for students and teachers. In R. Lesh & H. M. Doerr (Eds.), *Beyond Constructivism: Models and Modeling Perspectives on Mathematics Problem Solving, Learning, and Teaching* Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Newmann, F. M.; Bryk, A. S. & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: conflict or coexistence?* Chicago, IL: Consortium on Chicago School Research.

Newton, P. (2003). The defensibility of national curriculum assessment in England. *Research Papers in Education,* **18**(2), 101-127.

Nuthall, G. & Alton-Lee, A. (1995). Assessing classroom learning: how students use their knowledge and experience to answer classroom achievement test questions in science and social studies. *American Educational Research Journal,* **32**(1), 185-223.

Nyquist, J. B. (2003) *The benefits of reconstruing feedback as a larger system of formative assessment: a meta-analysis*. Unpublished Vanderbilt University Master of Science thesis.

Perrenoud, P. (1991). Towards a pragmatic approach to formative evaluation. In P. Weston (Ed.) *Assessment of pupil achievement* (pp. 79-101). Amsterdam, Netherlands: Swets & Zeitlinger.

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.

Reeves, J.; McCall, J. & MacGilchrist, B. (2001). Change leadership: planning, conceptualization and perception. In J. MacBeath & P. Mortimore (Eds.), *Improving school effectiveness* (pp. 122-137). Buckingham, UK: Open University Press.

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

Stokes, D. E. (1997). *Pasteur's quadrant: basic science and technological innovation.* Washington, DC: Brookings Institution Press.

White, B. Y. & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition. Making science accessible to all students. *Cognition and Instruction,* **16**(1), 3-118.

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. (1999, May) *A template for computer-aided diagnostic analyses of test outcome data*. Paper presented at 25th annual conference of the International Association for Educational Assessment held at Bled, Slovenia. London, UK: King's College London School of Education.

Wiliam, D. (2001). An overview of the relationship between assessment and the curriculum. In D. Scott (Ed.) *Curriculum and assessment* (pp. 165-181). Greenwich, CT: JAI Press.

Wiliam, D. (2003a). National curriculum assessment: how to make it better. *Research Papers in Education,* **18**(2), 129-136.

Wiliam, D. (2003b). The impact of educational research on mathematics education. In A. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Second International Handbook of Mathematics Education* (pp. 469-488). Dordrecht, Netherlands: Kluwer Academic Publishers.

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.

Wilson, S. M. & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: an examination of research on contemporary professional development. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (pp. 173-209). Washington, DC: American Educational Research Association.