Keeping learning on track:
formative assessment and the regulation of learning

Paper presented at the 20th biennial meeting of the Australian Association of Mathematics Teachers, Sydney, Australia, January 2005

Dylan Wiliam, ETS

Introduction

"I'd love to teach for deep understanding, but I have to raise my students' test scores." I have heard this sentiment from hundreds of teachers in many countries. Implicit in this statement is the notion that raising test scores is not compatible with teaching for deep understanding. As pressures for teachers to be accountable for the performance of their students increase, does this mean that there is no room for teaching for deep understanding? Or is there a way to achieve both?

Over the course of a 10-year study, Paul Black and I sought to find out if using assessment to *support* learning, rather than just to measure its results, can improve students' achievement, even when such achievement is measured in the form of state-mandated tests. In reviewing 250 studies from around the world, published between 1987 and 1998, we found that a focus by teachers on assessment *for* learning, as opposed to assessment *of* learning, produced a substantial increase in students' achievement (Black and Wiliam, 1998a). Since the studies also revealed that day-to-day classroom assessment was relatively rare, we felt that considerable improvements would result from supporting teachers in developing this aspect of their practice (Black and Wiliam, 1998b). The studies did not reveal, however, how this could be achieved and whether such gains would be sustained over an extended period of time.

Since 1999, we have worked with many groups of teachers, from both primary and secondary schools, in the United Kingdom and in the United States and these collaborations have shown that our initial optimism was justified. In a variety of settings teachers have found that teaching for deep understanding has resulted in an increase in student performance on externally-set tests and examinations (Wiliam et al, 2004).

The details of how we put these ideas into practice can be found elsewhere (Black et al, 2002; Black et al, 2003). In this paper, I want to describe the key ingredients of formative assessment: effective questioning, feedback, ensuring learners understand the criteria for success, and peer- and self-assessment, and then to show how they fit together within the general idea of the 'regulation of learning.'

What makes a good question?

Two items used in the Third International Mathematics and Science Study (TIMSS) are shown in figure 1 below. Although apparently quite similar, the success rates on the two items were very different. For example, in Israel, 88% of the students answered the first items correctly, while only 46% answered the second correctly, with 39% choosing response (b). The reason for this is that many students, in learning about fractions, develop the naive conception that the largest fraction is the one with the smallest denominator, and the smallest fraction is the one with the largest denominator. This approach leads to the correct answer for the first item, but leads to an incorrect response to the second. Furthermore, if we note that 46% plus 39% is very close to 88%, this provides strong evidence that many students who answered the first item correctly, did so with an incorrect strategy. In this sense, the first item is a much weaker item than the second, because many students can get it right for the wrong reasons.

This illustrates a very general principle in teachers' classroom questioning. By asking questions of students, teachers try to establish whether students have understood what they are meant to be learning, and if students answer the questions correctly, it is tempting to assume that the students' conceptions *match* those of the teacher. However, all that has really been established is that the students' conceptions *fit*, within the limitations of the questions. Unless the questions used are very rich, there will be a number of students who manage to give all the right responses, while having very different conceptions from those intended.

Item 1 (success rate 88%)

Which fraction is the smallest?

a) $\frac{1}{6}$ b) $\frac{2}{3}$ c) $\frac{1}{3}$ d) $\frac{1}{2}$

Item 2 (success rate 46%)

Which fraction is the largest?

a) $\frac{4}{5}$ b) $\frac{3}{4}$ c) $\frac{5}{8}$ d) $\frac{7}{10}$

Figure 1: two items from the Third International Mathematics and Science Study

A particularly stark example of this is the following pair of simultaneous equations:

$$\begin{aligned} 3a &= 24 \\ a + b &= 16 \end{aligned}$$

Many students find this difficult, saying that it can not be done. The teacher might conclude that they need some more help with equations of this sort, but the most likely reason for the difficulties with this item is not with mathematical skills but with their *beliefs*. If the students are encouraged to talk about their difficulty, they often say things like, "I keep on getting b is 8, but it can't be because a is." The reason that many students have developed such a belief is, of course, that before they were introduced to solving equations, they probably have been practicing substitution of numbers into algebraic formulas, where each letter stood for a different number. Although the students will not have been taught that each letter must stand for a different number, they have generalized implicit rules from their previous experience, just as because we always show them triangles where the lowest side is horizontal, they talk of "upside-down triangles" (Askew and Wiliam, 1995).

The important point here is that we would not have known about these unintended conceptions if the second equation had been a + b = 17 instead of a + b = 16. Items that reveal unintended conceptions—in other words that provide a “window into thinking”—are difficult to generate, but they are crucially important if we are to improve the quality of students’ mathematical learning.

Some people have argued that these unintended conceptions are the result of poor teaching. If only the teacher had phrased their explanation more carefully, had ensured that no unintended features were learned alongside the intended features, then these misconceptions would not arise.

But this argument fails to acknowledge two important points. The first is that this kind of over-generalization is a fundamental feature of human thinking. When young children say things like “I spended all my money,” they are demonstrating a remarkable feat of generalization. From the huge messiness of the language that they hear around them, they have learned that to create the past tense of a verb, one adds “d” or “ed.” In the same way, if one asks young children what causes the wind, the most common answer is “trees.” They have not been taught this, but have observed that trees are swaying when the wind is blowing and (like many politicians) have inferred a causation from a correlation.

The second point is that even if we wanted to, we are unable to control the student’s environment to the extent necessary for unintended conceptions not to arise. For example, it is well known that many students believe that the result of multiplying 2.3 by 10 is 2.30. It is highly unlikely that they have been taught this. Rather this belief arises as a result of observing regularities in what they see around them. The result of multiplying whole-numbers by 10 is just to add a zero, so why shouldn’t that work for all numbers? The only way to prevent students from acquiring this ‘misconception’ would be to introduce decimals before one introduces multiplying single-digit numbers by 10, which is clearly absurd. The important point is that we must acknowledge that what students learn is not necessarily what the teacher intended, and it is essential that teachers explore students’ thinking before assuming that students have ‘understood’ something. In this sense assessment is the bridge between teaching and learning.

Questions that give us this “window into thinking” are hard to find, but within any school there will be good selection of rich questions in use—the trouble is that each teacher will have her or his stock of good questions, but these questions don’t get shared within the school, and are certainly not seen as central to good teaching.

In most Anglophone countries, teachers spend the majority of their lesson preparation time in marking books, almost invariably doing so alone. In some other countries, the majority of lesson preparation time is spent planning how new topics can be introduced, which contexts and examples will be used, and so on. This is sometimes done individually or with groups of teachers working together. In Japan, however, teachers spend a substantial proportion of their lesson preparation time working together to devise questions to use in order to find out whether their teaching has been successful, in particular through the process known as ‘lesson study’ (Fernandez & Makoto, 2004).

Now in thinking up good questions, it is important not to allow the traditional concerns of reliability and validity to determine what makes a good question. For example, many teachers think that the following question, taken from the Chelsea Diagnostic Test for Algebra, is 'unfair':

Simplify (if possible): 2a + 5b

This item is felt to be unfair because students 'know' that in answering test questions, you have to do some work, so it must be possible to simplify this expression, otherwise the teacher wouldn't have asked the question. I would agree that to use this item in a test or an examination where the goal is to determine a student's achievement would probably not be a good idea. But to find out whether students understand algebra, it is a very good item indeed. If in the context of classroom work, rather than a formal test or exam, a student can be tempted to 'simplify' 2a + 5b then I want to know that, because it means that I have not managed to develop in the student a real sense of what algebra is about.

Similar issues are raised by asking students which of the following two fractions is the larger:

$$\frac{3}{7} \qquad \frac{3}{11}$$

Now in some senses this is a 'trick question.' There is no doubt that this is a very hard item, with typically only around one 14-year old in six able to give the correct answer (compared with around three-quarters of 14-year-olds being able to select correctly the larger of two 'ordinary' fractions). It may not, therefore, be a very good item to use in a test of students' achievement. But as a teacher, I think it is very important for me to know if my students think that $\frac{3}{11}$ is larger than $\frac{3}{7}$. The fact that this item is seen as a 'trick question' shows how deeply ingrained into our practice the summative function of assessment is.

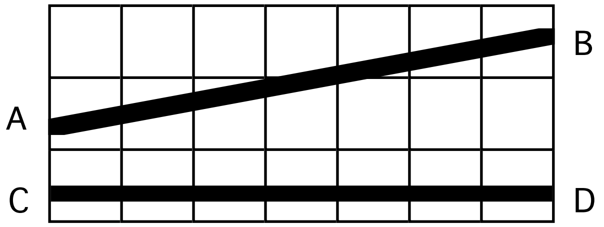
A third example, that caused considerable disquiet amongst teachers when it was used in a national test, is based on the following item, again taken from one of the Chelsea Diagnostic Tests:

Which of the following statements is true:

(1) AB is longer than CD

(2) AB is shorter than CD

(3) AB and CD are the same length



Again, viewed in terms of formal tests and examinations, this may be an unfair item, but in terms of a teacher's need to establish secure foundations for future learning, I would argue that this is entirely appropriate.

Rich questions, of the kind described above, provide teachers not just with evidence about what their students can do, but also what the teacher needs to do next, in order to broaden or deepen understanding.

Classroom questioning

There is also a substantial body of evidence about the most effective ways to use classroom questions. In many school, teachers tend to use questions are a way of directing the attention of the class, and keeping students 'on task' by scattering questions all around the classroom. This probably does keep the majority of students 'on their toes' but makes only a limited contribution to supporting learning. What is far less frequent is to see a teacher, in a whole-class lesson, have an extended exchange with a single student, involving a second, third, fourth or even fifth follow-up question to the student's initial answer. With such questions, the level of classroom dialogue can be built up to quite a sophisticated level, with consequent positive effects on learning. Of course, changing one's questioning style is very difficult where students are used to a particular set of practices (and may even regard asking supplementary questions as 'unfair'). And it may even be that other students see extended exchanges between the teacher and another student as a chance to relax and go 'off task' but as soon as students understand that the teacher may well be asking them what they have learned from a particular exchange between another student and the teacher, their concentration is likely to be quite high.

How much time a teacher allows a student to respond before evaluating the response is also important. It is well known that teachers do not allow students much time to answer questions, and, if they don't receive a response quickly, they will 'help' the student by providing a clue or weakening the question in some way, or even moving on to another student. However, what is not widely appreciated is that the amount of time between the student providing an answer and the teacher's evaluation of that answer is just as, if not more, important. Of course, where the question is a simple matter of factual recall, then allowing a student time to reflect and expand upon the answer is unlikely to help much. But where the question requires thought, then increasing the time between the end of the student's answer and the teacher's evaluation from the average 'wait-time' of less than a second to three seconds, produces measurable increases in learning (although increases beyond three seconds have little effect, and may cause lessons to lose pace).

In fact, questions need not always come from the teacher. There is substantial evidence that students' learning is enhanced by getting them to generate their own questions Foos et al, 1994). If instead of writing an end-of-topic test herself, the teacher asks the students to write a test that tests the work the class has been doing, the teacher can gather useful evidence about what the students think they have been learning, which is often very different from what the teacher thinks the class has been learning. This can be a particularly effective strategy with disaffected older students, who often feel threatened by tests. Asking them to write a test for the topic they have completed, and making clear that the teacher is going to mark the question rather than the answers, can be a hugely liberating experience for many students.

Some researchers have gone even further, and shown that questions can limit classroom discourse, since they tend to demand a simple answer. There is a substantial body of evidence the classroom learning is enhanced considerably by shifting from asking questions to making statements (Dillon, 1988). For example, instead of asking "Are all squares rectangles", which seems to require a 'simple' yes/no answer, the level of classroom discourse (and student learning) is improved considerably by framing the same question as a statement—"All squares are rectangles", and asking students to discuss this in small groups before presenting a reasoned conclusion to the class.

The quality of feedback

Ruth Butler (1998) investigated the effectiveness of different kinds of feedback on 132 year 7 students in 12 classes in 4 Israeli schools. For the first lesson, the students in each class were given a booklet containing a range of divergent thinking tasks. At the end of the lesson, their work was collected in. This work was then marked by independent markers. At the beginning of the next lesson, two days later, the students were given feedback on the work they had done in the first lesson. In four of the classes students were given marks (which were scaled so as to range from 40 to 99) while in another four of the classes, students were given comments, such as "You thought of quite a few interesting ideas; maybe you could think of more ideas." In the other four classes, the students were given both marks and comments.

Then, the students were asked to attempt some similar tasks, and told that they would get the same sort of feedback as they had received for the first lesson's work. Again, the work was collected in and marked.

Those given only marks made no gain from the first lesson to the second. Those who had received high marks in the tests were interested in the work, but those who had received low marks were not. The students given only comments scored, on average, 30% more on the work done in the second lesson than on the first, and the interest of all the students in the work was high. However, those given both marks and comments made *no gain* from the first lesson to the second, and those who had received high marks showed high interest while those who received low marks did not.

In other words, far from producing the best effects of both kinds of feedback, giving marks alongside the comments completely washed out the beneficial effects of the comments. The use of both marks and comments is probably the most widespread form of feedback used in the Anglophone world, and yet this study (and others like it—see below) show that it is no more effective than marks alone. In other words, if you write careful diagnostic comments on a student's work, and then put a score or grade on it, you are wasting your time. The students who get the high scores do not need to read the comments and the students who get the low scores do not want to. You would be better off just giving a score. The students will not learn anything from this but you will save yourself a great deal of time.

A clear indication of the role that ego plays in learning is given by another study by Ruth Butler (1987). In this study, 200 year 6 and 7 students spent a lesson working on a variety of divergent thinking tasks. Again, the work was collected in and the students were given one of four kinds of feedback on this work at the beginning of the second lesson (again two days later):

a quarter of the students were given comments;
a quarter were given grades;

a quarter were given written praise; and
a quarter were given no feedback at all.

The quality of the work done in the second lesson was compared to that done in the first. The quality of work of those given comments had improved substantially compared to the first lesson, but those given grades and praise had made no more progress than those given absolutely no feedback throughout their learning of this topic.

At the end of the second lesson, the students were given a questionnaire about what factors influenced their work. In particular the questionnaire sought to establish whether the students attributed successes and failures to themselves (called ego-involvement) or to the work they were doing (task-involvement). Examples of ego- and task-involving attributions are shown in table 1.

Attribution of	Ego	Task
Effort	To do better than others To avoid doing worse than others	Interest To improve performance
Success	Ability Performance of others	Interest Effort Experience of previous learning

Table 1: ego- and task-related attributions

Those students given comments during their work on the topic had high levels of task-involvement, but their levels of ego-involvement were the same as those given no feedback. However, those given praise and those given grades had comparable levels of task-involvement to the control group, but their levels of ego-involvement were substantially higher. The only effect of the grades and the praise, therefore, was to increase the sense of ego-involvement without increasing achievement.

This should not surprise us. In pastoral work, we have known for many years that one should criticize the behaviour, not the child, thus focusing on task-involving rather than ego-involving feedback. These findings are also consistent with the research on praise carried out in the 1970s which showed clearly that praise was not necessarily ‘a good thing’—in fact the best teachers appear to praise slightly less than average (Good and Grouws, 1975). It is the quality, rather than the quantity of praise that is important and in particular, teacher praise is far more effective if it is infrequent, credible, contingent, specific, and genuine (Brophy, 1981). It is also essential that praise is related to factors within an individual’s control, so that praising a gifted student just for being gifted is likely to lead to negative consequences in the long term.

The timing of feedback is also crucial. If it is given too early, before students have had a chance to work on a problem, then they will learn less. Most of this research has been done in the United States, where it goes under the name of ‘peekability research’, because the important question is whether students are able to ‘peek’ at the answers before they have tried to answer the question. However, a British study, undertaken by Simmonds and Cope (1993) found similar results. Pairs of students aged between 9 and 11 worked on angle and rotation problems. Some of these worked on the problems using Logo and some worked on the problems using pencil and paper. The students working in Logo were able to use a ‘trial and improvement’ strategy that enabled them to get a solution with little mental effort. However, for those working with pencil and paper, working out the effect of a single rotation was much more time consuming, and thus the students had an incentive to think carefully, and this greater ‘mindfulness’ led to more learning.

The effects of feedback highlighted above might suggest that the more feedback, the better, but this is not necessarily the case. Day and Cordon (1993) looked at the learning of a group of 64 year 4 students on reasoning tasks. Half of the students were given a ‘scaffolded’ response when they got stuck—in other words they were given only as much help as they needed to make progress, while the other half were given a complete solution as soon as they got stuck, and then given a new problem to work on. Those given the ‘scaffolded’ response learned more, and retained their learning longer than those given full solutions.

In a sense, this is hardly surprising, since those given the complete solutions had the opportunity for learning taken away from them. As well as saving time, therefore, developing skills of ‘minimal intervention’ promote better learning.

Sometimes, the help need not even be related to the subject matter. Often, when a student is given a new task, the student asks for help immediately. When the teacher asks, “What can’t you do?” it is common to hear the reply, “I can’t do any of it.” In such circumstances, the student’s reaction may be caused by anxiety about the unfamiliar nature of the task, and it is frequently possible to support the student by saying something like “Copy out that table, and I’ll be back in five minutes to help you fill it in.” This is often all the support the student needs. Copying out the table forces the student to look in detail at how the table is laid out, and this ‘busy work’ can provide time for the student to make sense of the task herself.

The consistency of these messages from research on the effects of feedback extends well beyond school and other educational settings. A review of 131 well-designed studies in educational and workplace settings found that, on average, feedback did improve performance, but this average effect disguised substantial differences between studies. Perhaps most surprisingly, in 40% of the studies, giving feedback had a negative impact on performance. In other words, in two out of every five carefully-controlled scientific studies, giving people feedback on their performance made their performance worse than if they were given no feedback on their performance at all! On further investigation, the researchers found that feedback makes performance worse when it is focused on the self-esteem or self-image (as is the case with grades and praise). The use of praise can increase motivation, but then it becomes necessary to use praise all the time to maintain the motivation. In this situation, it is very difficult to maintain praise as genuine and sincere. In contrast, the use of feedback improves performance when it is focused on what needs to be done to improve, and particularly when it gives specific details about *how* to improve.

This suggests that feedback is not the same as formative assessment. Feedback is a necessary first step, but feedback is formative *only if the information fed back to the learner is used by the learner in improving performance*. If the information fed back to the learner is intended to be helpful, but cannot be used by the learner in improving her own performance it is not formative. It is rather like telling an unsuccessful comedian to “be funnier.”

As noted above, the quality of feedback is a powerful influence on the way that learners attribute their successes and failures. A series of research studies, carried out by Carol

Dweck over twenty years (see Dweck, 2000 for a summary), has shown that different students differ in the whether they regard their success and failures as:

being due to ‘internal’ factors (such as one’s own performance) or ‘external’ factors (such as getting a lenient or severe marker);

being due to ‘stable’ factors (such as one’s ability) or ‘unstable’ factors (such as effort or luck); and

applying globally to everything one undertakes, or related only to the specific activity on which one succeeded or failed.

Table 2 gives some examples of attributions of success and failure.

Attribution	Success	Failure
locus	internal: “I got a good mark because it was a good piece of work”	internal: “I got a low mark because it wasn’t a very good piece of work”
	external: “I got a good mark because the teacher likes me”	external: “I got a low mark because the teacher doesn’t like me”
stability	stable: “I got a good exam-mark because I’m good at that subject”	stable: “I got a bad exam-mark because I’m no good at that subject”
	unstable: “I got a good exam-mark because I was lucky in the questions that came up”	unstable: “I got a bad exam-mark because I hadn’t done any revision”
specificity	specific: “I’m good at that but that’s the only thing I’m good at”	specific: “I’m no good at that but I’m good at everything else”
	global: “I’m good at that means I’ll be good at everything”	global: “I’m useless at everything”

Table 2: dimensions of attributions of success and failure

Dweck and others have found that boys are more likely to attribute their successes to stable causes (such as ability), and their failures to unstable causes (such as lack of effort and bad luck). This would certainly explain the high degree of confidence with which boys approach tests or examinations for which they are completely unprepared. More controversially, the same research suggests that girls attribute their successes to unstable causes (such as effort) and their failures to stable causes (such as lack of ability), leading to what has been termed ‘learned helplessness.’

More recent work in this area suggests that what matters more, in terms of motivation, is whether students see ability as fixed or incremental. Students who believe that ability is fixed will see any piece of work that they are given as a chance either to re-affirm their ability, or to be ‘shown-up.’ If they are confident in their ability to achieve what is asked of them, then they will attempt the task. However, if their confidence in their ability to carry out their task is low, then, unless the task is so hard that no-one is expected to succeed, they will avoid the challenge, and this can be seen in mathematics classrooms all over the world every day. Taking all things into account, a large number of students decide that they would rather be thought lazy than stupid, and refuse to engage with the task, and this is a direct consequence of the belief that ability is fixed. In contrast, those who see ability as incremental see all challenges as chances to learn—to get cleverer—and therefore in the face of failure will try harder. What is perhaps most important here is that these views of ability are generally not global—the same students often believe that ability in schoolwork is fixed, while at the same time believe that ability in athletics is incremental, in that the more one trains, the more one's ability increases. What we therefore need to do is to ensure that the feedback we give students supports a view of ability as incremental rather than fixed.

Perhaps surprisingly for educational research, the research on feedback paints a remarkably coherent picture. Feedback to learners should focus on what they need to do to improve, rather than on how well they have done, and should avoid comparison with others. Students who are used to having every piece of work scored or graded will resist this, wanting to know whether a particular piece of work is good or not, and in some cases, depending on the situation, the teacher may need to go along with this. In the long term, however, we should aim to reduce the amount of ego-involving feedback we give to learners (and with new entrants to the school, not begin the process at all), and focus on the student's learning needs. Furthermore, feedback should not just tell students to work harder or be 'more systematic'—the feedback should contain a recipe for future action, otherwise it is not formative. Finally, feedback should be designed so as to lead all students to believe that ability—even in mathematics—is incremental. In other words the more we 'train' at mathematics, the clever we get.

Although there is a clear set of priorities for the development of feedback, there is no 'one right way' to do this. The feedback routines in each class will need to be thoroughly integrated into the daily work of the class, and so it will look slightly different in every classroom. This means that no one can tell teachers how this should be done—it will be a matter for each teacher to work out a way of incorporating some of these ideas into her or his own practice.

Sharing criteria with learners

Frederiksen and White (1997) undertook a study of three teachers, each of whom taught 4 parallel year 8 classes in two U.S. schools. The average size of the classes was thirty-one. In order to assess the representativeness of the sample, all the students in the study were given a basic skills test, and their scores were close to the national average. All twelve classes followed a novel curriculum (called ThinkerTools) for a term. The curriculum had been designed to promote thinking in the science classroom through a focus on a series of seven scientific investigations (approximately two weeks each). Each investigation incorporated a series of evaluation activities. In two of each teacher's four classes these evaluation episodes took the form of a discussion about what they liked and disliked about the topic. For the other two classes they engaged in a process of 'reflective assessment.' Through a series of small-group and individual activities, the students were introduced to the nine assessment criteria (each of which was assessed on a 5-point scale) that the teacher would use in evaluating their work. At the end of each episode within an investigation, the students were asked to assess their performance against two of the criteria, and at the end of the investigation, students had to assess their performance against all nine. Whenever they assessed themselves, they had to write a brief statement showing which aspects of their work formed the basis for their rating. At the end of each investigation, students presented their work to the class, and the students used the criteria to give each other feedback.

As well as the students' self-evaluations, the teachers also assessed each investigation, scoring both the quality of the presentation and the quality of the written report, each being scored on a 1 to 5 scale. The possible score on each of the seven investigations therefore ranged from 2 to 10.

The mean project scores achieved by the students in the two groups over the seven investigations are summarized in table 3, classified according to their score on the basic skills test.

	Score on basic skills test		
Group	Low	Intermediate	High
Likes and dislikes	4.6	5.9	6.6
Reflective assessment	6.7	7.2	7.4

Note: the 95% confidence interval for each of these means is approximately 0.5 either side of the mean

Table 3: Mean project scores for students

Two features are immediately apparent in these data. The first is that the mean scores are higher for the students doing 'reflective assessment,' when compared with the control group—in other words, all students improved their scores when they thought about what it was that counted as good work. However, much more significantly, the difference between the 'likes and dislikes' group and the 'assessment' group was much greater for students with weak basic skills. This suggests that, at least in part, low achievement in schools is exacerbated by students' not understanding what it is they are meant to be doing—an interpretation borne out by the work of Eddie Gray and David Tall (1994), who have shown that 'low-attainers' often struggle because what they are trying to do is actually much harder than what the 'high-attainers' are doing. This study, and others like it, shows how important it is to ensure that students understand the criteria against which their work will be assessed. Otherwise we are in danger of producing students who do not understand what is important and what is not. As the old joke about project work has it: "four weeks on the cover and two on the contents."

Now although it is clear that students need to understand the standards against which their work will be assessed, the study by Frederiksen and White shows that the criteria themselves are only the starting point. At the beginning, the words do not have the meaning for the student that they have for the teacher. Just giving 'quality criteria' or 'success criteria' to students will not work, unless students have a chance to see what this might mean in the context of their own work.

Because we understand the meanings of the criteria that we work with, it is tempting to think of them as *definitions* of quality, but in truth, they are more like labels we use to talk about ideas in our heads. For example, 'being systematic' in an investigation is not something we can define explicitly, but we can help students develop what Guy Claxton calls a 'nose for quality.'

One of the easiest ways of doing this is to do what Frederiksen and White did. Marking schemes are shared with students, but they are given time to think through, in discussion with others, what this might mean in practice, applied to their own work. We shouldn't assume that the students will understand these right away, but the criteria will provide a focus for negotiating with students about what counts as quality in the mathematics classroom

Another way of helping students understand the criteria for success is, before asking the students to embark on (say) an investigation, to get them to look at the work of other students (suitably anonymized) on similar (although not, of course the same) investigations. In small groups, they can then be asked to decide which of pieces of students' work are good investigations, and why. It is not necessary, or even desirable, for the students to come to firm conclusions and a definition of quality—what is crucial is that they have an opportunity to explore notions of 'quality' for themselves. Spending time looking at other students' work, rather than producing their own work, may seem like 'time off-task', but the evidence is that it is a considerable benefit, particularly for 'low-attainers.'

Student peer- and self-assessment

Whether students can really assess their own performance objectively is a matter of heated debate, but very often the debate takes place at cross-purposes. Opponents of self-assessment say that students cannot possibly assess their own performance objectively, but this is an argument about *summative* self-assessment; no-one is seriously suggesting that students ought to be able to write their own school-leaving certificates. What really matters is whether self-assessment can enhance learning, and in this regard, accuracy is a secondary concern.

The power of student self-assessment is shown very clearly in an experiment by Fontana and Fernandez (1994). A group of 25 Portuguese primary school teachers met for two hours each week over a twenty-week period during which they were trained in the use of a structured approach to student self-assessment. The approach to self-assessment involved an exploratory component and a prescriptive component. In the exploratory component, each day, at a set time, students organized and carried out individual plans of work, choosing tasks from a range offered to them by the teacher, and had to evaluate their performance against their plans once each week. The progression within the exploratory component had two strands—over the twenty weeks, the tasks and areas in which the students worked were to take on the student's own ideas more and more, and secondly, the criteria that the students used to assess themselves were to become more objective and precise.

The prescriptive component took the form of a series of activities, organized hierarchically, with the choice of activity made by the teacher on the basis of diagnostic assessments of the students. During the first two weeks, children chose from a set of carefully structured tasks, and were then asked to assess themselves. For the next four weeks, students constructed their own mathematical problems following the patterns of those used in weeks 1 and 2, and evaluated them as before, but were required to identify any problems they had, and whether they had sought appropriate help from the teacher.

Over the next four weeks, students were given further sets of learning objectives by the teacher, and again had to devise problems, but now, they were not given examples by the teacher. Finally, in the last ten weeks, students were allowed to set their own learning objectives, to construct relevant mathematical problems, to select appropriate apparatus, and to identify suitable self-assessments.

Another 20 teachers, matched in terms of age, qualifications, experience, using the same curriculum scheme, for the same amount of time, and doing the same amount of inservice training, acted as a control group. The 354 students being taught by the 25 teachers using self-assessment, and the 313 students being taught by the 20 teachers acting as a control group were each given the same mathematics test at the beginning of the project, and again at the end of the project. Over the course of the experiment, the marks of the students taught by the control-group teachers improved by 7.8 marks. The marks of the students taught by the teachers developing self-assessment improved by 15 marks—almost twice as big an improvement.

Now the details of the particular approach to self-assessment are not given in the paper, and are in any case not that important—Portuguese primary schools are, after all, very different from those in other countries. However this is just one of a huge range of studies, in different countries, and looking at students of different ages, that have found a similar pattern. Involving students in assessing their own learning improves that learning.

The regulation of learning

Although at first sight quite different, the four elements of effective formative assessment outlined above form a coherent set of strategies for raising achievement. The coherence of these ideas can be seen more clearly by considering three crucial processes in learning:

where the learners are in their learning
where they are going,
how to get there,

and the role of the learner, her or his peers, and the teacher in these processes. The result of crossing these two dimensions is shown in table 4.

	Where the learner is	Where they are going	How to get there
Teacher	Evoking information	Establishing goals	Feedback
Peer	Peer-assessment	Sharing success criteria	Peer-tutoring
Student	Self-assessment	Sharing success criteria	Self-directed learning

Table 4: aspects of formative assessment

Rich questioning and effective feedback focus on the teacher's role—first being clear about where we want students to get to (curricular goals), asking appropriate questions to find out where they are, and feeding back to students in ways that the students can use in improving their own performance. Sharing criteria with learners and student self-assessment focus on the learner's role—first being clear about where they want to get to, and then monitoring their own progress towards that goal.

The elements in table 4 can be integrated within a more general theoretical framework of the *regulation of learning processes* as suggested Perrenoud (1991, 1998)[1]. Within such a

[1] 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

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.

From this perspective, the task of the teacher is not necessarily to teach, but to create situations in which students learn. This focus emphasizes what it is that students learn, rather than what teachers do. Most teachers appear to be quite skilled at regulating or controlling the activities in which students engage, 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 away from regulating the activities in which students engage, 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.

However, in this context, it is important to note that the 'engineering of learning environments' does not guarantee that the learning is proceeds in fruitful ways. 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. An emphasis on the regulation of learning processes entails ensuring that the learning that is taking place is as intended.

When the learning environment is well-regulated, much of the regulation is pro-active, 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 organization of regulation of learning processes run by technologies or incorporated into classroom organization and management" (Perrenoud, 1998 p100). For example, a teacher's decision to use realistic contexts in the mathematics classroom can provide a source of proactive 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?" or "Would your mass be the same on the moon?" (In this context it is 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.)

unfortunate connotation of the second. In French, the two senses have separate terms (régulation and règlement) and so the problem does not arise.

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 constantly have 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, 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 goal of learning is generally both highly specific and common to all the students in a class. In contrast, when the class is doing an investigation, 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 or the United Kingdom, 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.

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 pro-active or ‘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.

Summary

In this paper, I have outlined some of the research that suggests that focusing on the use of day-to-day formative assessment is one of the most powerful ways of improving learning in the mathematics classroom. In other words, even if teachers do not care about deep understanding, and instead wish only to increase their students' test scores, then attention to formative assessment appears to be one of, if not the, most powerful way to do this.

To be effective, these strategies must be embedded into the day-to-day life of the classroom, and must be integrated into whatever curriculum scheme is being used. That is why there can be no recipe that will work for everyone. Each teacher will have to find a way of incorporating these ideas into their own practice, and effective formative assessment will look very different in different classrooms. It will, however, have some distinguishing features. Students will be thinking more often than they are trying to remember something, they will believe that by working hard, they get cleverer, they will understand what they are working towards, and will know how they are progressing.

In some ways, this is an old-fashioned message—indeed, none of the strategies that teachers have used to put these principles into practice in their classrooms is new. What is new is that we now have hard empirical evidence that quality learning does lead to higher achievement, even when performance is measured through externally-mandated tests. What is also new is the broad theoretical framework of the regulation of learning, which may help teachers to understand how these ideas can be implemented effectively, so that teachers and students can, together, keep the learning of mathematics 'on –track'

References

Askew, M. & Wiliam, D. (1995). *Recent research in mathematics education 5-16*. London, UK: Her Majesty's Stationery Office.

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

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

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

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

Butler, R. (1987) Task-involving and ego-involving properties of evaluation: effects of different feedback conditions on motivational perceptions, interest and performance. *Journal of Educational Psychology* **79** (4) 474-482.

Butler, R. (1988) Enhancing and undermining Intrinsic motivation; the effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology* **58** 1-14.

Dillon, J. T. (1988). *Questioning and teaching: a manual of practice*. London: Croom Helm.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist,* **41**(10), 1040-1048.

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

Fernandez, C. & Makoto, Y. (2004). *Lesson study: a Japanese approach to improving mathematics teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fontana, D. & Fernandes, M. (1994). Improvements in mathematics performance as a consequence of self-assessment in Portuguese primary school pupils. *British Journal of Educational Psychology,* **64**, 407-417.

Foos, P. W.; Mora, J. & Tkacz, S. (1994). Student study techniques and the generation effect. *Journal of Educational Psychology,* **86**(4), 567-576.

Frederiksen, J. R. & White, B. Y. (1997). Reflective assessment of students' research within an inquiry-based middle school science curriculum. Paper presented at the *Annual meeting of the American Educational Research Association.* Chicago, IL.

Good, T. L. and Grouws, D. A. (1975). Process-product relationships in fourth grade mathematics classrooms. Report for National Institute of Education, Columbia, MO: University of Missouri (report no NE-G-00-0-0123).

Gray, E. M. & Tall, D. O. (1994). Duality, ambiguity and flexibility: a 'proceptual' view of simple arithmetic. *Journal for Research in Mathematics Education,* **25**(2), 116-140.

Hart, K. M.; Brown, M. L.; Kerslake, D.; Küchemann, D. & Ruddock, G. (1985). *Chelsea diagnostic mathematics tests.* Windsor, UK: NFER-Nelson.

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.

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.

Dylan Wiliam is director of the Learning and Teaching Research Center at the Educational Testing Service. He can be reached at dylanwiliam@mac.com.