Chapter 1

Improving Learning in Science With Formative Assessment

Dylan Wiliam, Institute of Education, University of London

[A-Head] Introduction

In recent years, the No Child Left Behind law has focused attention on student achievement in science across the United States, but there are more important reasons for being concerned with student achievement. Increasing student achievement brings substantial benefits both to the individual and to society. For the individual, improved school achievement increases career earnings, improves health and well-being, and actually lengthens life (Wiliam and Thompson 2007; Lleras-Muney 2005). For society, the benefits include increased tax revenues, savings in public health costs, reduced law-enforcement and prison costs, as well as savings in welfare budgets (Hoff 2007).

Achievement in science in particular is likely to be increasingly important in the future for the needs of employment, but it will also be essential for democratic citizenship. Without an understanding of what science can do (and what it can't) and how science does what it does, public policies about issues such as genetically modified foods, assisted reproductive technologies, and human cloning are likely to be set on the basis of populist journalism rather than scientific evidence.

The focus of this chapter, therefore, is about how we can raise achievement in science in the United States. In this chapter I will argue that if we are serious about raising achievement in science, then we need to look beyond "what works" in education to notions of cost-benefit—not just whether a particular initiative raises achievement, but by how much, and at what cost. I will show that the evidence currently at hand suggests that this is done by investing in teacher professional development, but of a sort very different from what occurs in most school districts currently. I will also show that this professional development needs to be focused specifically on changing what teachers do in the classroom and, in particular, needs to be aimed at changing teachers' minute-by-minute and day-by-day use of assessment to modify instruction, sometimes called formative assessment or assessment for learning.

[A-Head] Value for Money in Education Reform

Many policy makers have focused on "what works" in education, but such a focus is misguided, since what is most is important is the size of effect relative to cost. An intervention might "work" but the effects might be too small to be worth bothering with, or it might produce substantial effects but be too expensive to implement. The focus in school improvement needs to be on the ratio of the size of the benefit to the cost incurred in bringing it about.

When we adopt this perspective, we find that some effective interventions are too costly, and some interventions that have only a small impact on student learning nevertheless

turn out to be implementable at modest cost. To take one example, the research shows clearly that reducing class size raises student achievement. So what? The important question should be how much improvement, and at what cost? And here the data are depressing.

Jepsen and Rivkin (2002) found that reducing elementary school class size by 10 students would increase the proportion of students passing typical mathematics and reading tests by 4% and 3% respectively. While there is evidence that the effects are larger for the early years, beyond this point, class-size reduction appears to be a very expensive way of increasing student achievement. To make this more precise, consider the cost of this intervention per classroom of 30 students. Reducing class-size from (say) 30 to 20 would increase the salary costs by approximately 50%, because we would need three teachers instead of two for a group of 60 students. Assuming an average teacher salary cost of $60,000 per year, this would cost $30,000 per year to implement.

Using tests such as those used in international comparisons such as TIMSS (Rodriguez 2004) as a benchmark, we find that the effect of these class-size reductions would be equivalent to students in the smaller classes learning in three months what it would take the students in the larger classes four months to learn; a 33% increase in the speed of learning.

A more recent study by Jenkins, Lovacic, and Vignoles (2006) in England found that additional resources, if used to reduce class size, might have a larger increase on student achievement in middle and high school science (up to four times the effect size found by Jepsen and Rivkin), but this model assumed that additional science teachers of equivalent quality to those already on the job could be found, which is at the very least questionable (one of the key factors in the modest results in the Jepsen and Rivkin research cited above was that the extra teachers were not as good as those already in place). Nevertheless, it may be that there are special cases, such as early years, and secondary science, where class-size reduction may have a substantial effect on student achievement. However, in general, class-size reduction would appear to offer only modest increases in student achievement at very high cost.

How else could we obtain the same increase in student achievement? One obvious candidate is to increase teacher subject knowledge. Hill, Rowan, and Ball (2005) found that an increase of two standard deviations in what they termed “mathematical knowledge for teaching” was associated with an increase of up to 0.1 standard deviations in student achievement in mathematics. This was not a direct experiment, and so we cannot infer a causal relationship. However, it suggests that the same improvement in achievement that would be gained by reducing class size from 30 to 20 might be secured by increasing teacher subject-matter knowledge by two standard deviations. Unfortunately, there is currently little evidence about how to do this.

In contrast, supporting teachers in developing the use of assessment for learning has been shown to roughly double the speed of learning (Wiliam et al. 2004; Hayes 2003). In other words, students learned in six months what would have taken a year to learn in other classrooms. The cost of this intervention was around $8,000 per teacher, but of course,

unlike class-size reduction, the cost of which has to be found annually, investing in teachers' professional development is a one-off expenditure, which can be depreciated over the teacher's remaining career. Against this, some cost of annual renewal needs to be allowed. Assuming that a new teacher will continue to teach for at least five years, and allowing time for four hours of meetings with colleagues per month for renewal, the cost of teacher professional development focused on assessment for learning would appear to be around $3,000 per teacher (and therefore per classroom) per year.

Compared to class-size reduction, therefore, improving teachers' use of assessment for learning would appear to promise two or three times the increase in student learning, for around one-tenth the cost. Even in the special case of secondary school science discussed above, assessment for learning produces the roughly the same size of benefit as reducing class size by 30%, at less than one-fifth the cost.

[A-Head] Assessment for Learning

When teachers are asked how they assess their students, they typically talk about tests, examinations, quizzes, and other formal methods. When they are asked how they know whether their students have learned what they have taught, the answers are very different. They talk about homework, classwork, the things students say in classroom discussions, and even the expressions on their faces. This is the great disconnect in education worldwide. Assessment that serves the needs of teachers and their students is seen as completely separate from, and indeed, incompatible with, assessment that serves the needs of parents, administrators, policy makers, and other stakeholders.

At one extreme we have a teacher questioning a student, trying to elicit evidence of (mis)conceptions that are likely to impede future learning. At the other extreme we have the use of Advanced Placement tests, used both to give students college-level credit and, by some universities, to decide which students to admit. The obvious conclusion is that the latter kind of assessment hinders learning while the former helps, but things are not that simple.

The teacher's questioning of the student can cause damage, possibly irreparable, to the student's sense of self if undertaken in a humiliating way. And at the other extreme, Advanced Placement tests provide clear guidance to teachers and students about what the vague words in examination syllabuses mean. Furthermore, when used as "trial" examinations, sample papers allow students to benchmark themselves against the standard established by the College Board and to help each other rectify deficiencies. If we are to design assessment systems that help rather than hinder learning, we must go beyond looking at the assessments themselves and look at deeper issues about how the assessments help learners and their teachers know where the learners are in their learning, where they are going, and how to get there.

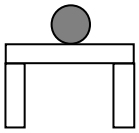
Through extensive reviews of the available research evidence, and through extensive field work with teachers, both in the UK and in the United States (see the suggestions for further reading at the end of this chapter), we have identified five "key strategies," which, when implemented appropriately, allow assessment to help, rather than hinder, learning.

1. Engineer effective classroom discussions, questions, activities, and tasks that elicit evidence of student learning.

The first step in using assessment to help learning is to collect the right sort of evidence, and here it is clear that the tools that teachers use to find out where students are in their learning are given too little attention.

Few teachers plan the kinds of tasks, activities, and questions that they use with their students specifically to elicit the right kind of evidence of student learning. As an example of a good question, consider the science question shown in Figure 1.1.

Figure 1.1 Science Item



The ball resting on the table is not moving. It is not moving because:

A. no forces are pushing or pulling on the ball.

B. gravity is pulling down, but the table is in the way.

C. the table pushes up with the same force that gravity pulls down

D. gravity is holding it onto the table.

E. there is a force inside the ball keeping it from rolling off the table

Source: Adapted by the author from Wilson, M., and K. Draney. 2004. Some links between large-scale and classroom assessments: The case of the BEAR assessment system. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education.* Chicago, IL: University of Chicago Press.

Response A clearly relates to the well-known misconception that if there is no movement, then there are no forces acting. Response B is more complex. In one sense there is nothing wrong with B. After all, gravity *is* pulling the ball down, and the table *is* in the way. The reason that B is a less preferable response to C is not that it is not correct; it is that it is not physics. Science provides us with a powerful, but unnatural, way of thinking about the world; after all, if it were natural, we would not need to teach it. Students who choose B rather than C have not understood the important idea that lack of movement denotes forces in equilibrium, not the absence of forces.

By crafting questions that explicitly build in the under- and over-generalizations that we know students make, we can get far more useful information about what to do next. By

equipping each student in the class with a set of cards with A, B, C, D, and E on them, and by requiring all students to respond simultaneously with their answers, the teacher can generate a very solid evidence base for deciding whether the class is ready to move on. If every student responds with C, then the teacher can move on with confidence that the students have understood. If everyone simply responds with A, then the teacher may choose to re-teach some part of the topic. The most likely response, however, is for some students to respond correctly and for others to respond incorrectly. This provides the teacher with an opportunity to conduct a classroom discussion in which students with different views can be asked to justify their selections. Moreover, because the teacher knows which student gave which response, the discussion can be better orchestrated (e.g., "Shane, you also chose B. Did you choose it for the same reason that Alicia gave, or a different reason?").

Of course, planning such questions takes time, but by investing the time before the lesson, teachers are able to address students' confusion during the lesson, with the students still in front of them. Teachers who do not plan such questions are forced to put children's thinking back on track by giving extended comments while grading, thus dealing with the students one at a time, after they have left the classroom.

2. Provide feedback that moves learning forward

The research on feedback shows that much of the feedback that students receive has, at best, no impact on learning, and can actually be counterproductive. One extraordinary study (Kluger and DeNisi 1996) reviewed over 3,000 research reports on the effects of feedback in schools, colleges, and workplaces. They then rejected studies that failed to reach the highest standards of methodological rigor and were left with just 131 studies. Across these 131 studies, they found that, on average, feedback did increase achievement, but that in 50 of the studies (i.e., almost two in five), feedback actually made people's performance worse than it would have been without feedback. The key feature of these studies was that feedback was, in the psychological jargon, "ego-involving." In other words, the feedback focused attention on the person rather than the quality of the work—for example, by giving scores, grades, or other forms of report that encouraged comparison with others. For the 81 studies that found a positive impact on performance, Kluger and DeNisi found that the biggest impacts occurred when feedback told not just what to do to improve, but also how to go about it.

An example from athletics may be helpful here. If a young fast-pitch softballer has an ERA (earned run average) of 10, we know that she is not doing well. This is the *monitoring* assessment. The monitoring assessment identifies that there is a problem, but doesn't identify what it is. By looking at her pitching, we might see that the reason that her ERA is so high is that she is trying, unsuccessfully, to pitch a rising fastball. This is a pitch that is thrown with enough backspin so that as it reaches the batter, it rises sharply, making it almost unhittable. The problem is that if it does not rise, then the result is a fastball over the center of the plate, which is very easy to hit, and this is exactly what is happening to our pitcher. Her rising fastball is not rising. This is the *diagnostic* assessment. The diagnostic assessment identifies where the problem is, but by itself, doesn't give the athlete any clue about how to go about making improvements. However,

if the pitching coach can see that the reason that the pitcher is struggling to pitch the rising fastball is because she is not dropping the pitching shoulder low enough to deliver the pitch from below the knee, then this gives the athlete something to work with. This is the *formative* assessment. Just as we use the term *formative* to describe the experiences that shape us as we grow up, a formative assessment is one that shapes learning. Much of the feedback that students get while learning science is no more helpful than telling our fast-pitch softballer to make sure her rising fastball rises or telling a bad comedian that he needs to be funnier. If feedback is to impact learning it must focus on what needs to happen next; in other words, it must be a *guide* to action and not just a demand for it. For an example of a feedback system focused on middle school science, see Clymer and Wiliam (2006/2007).

3. Clarifying and sharing learning intentions and success criteria with learners

In an article entitled "The View From the Student's Desk," written over 35 years ago, Mary Alice White (1971) said:

BLOCK QUOTE

> The analogy that might make the student's view more comprehensible to adults is to imagine oneself on a ship sailing across an unknown sea, to an unknown destination. An adult would be desperate to know where he [sic] is going. But a child only knows he is going to school.... The chart is neither available nor understandable to him... Very quickly, the daily life on board ship becomes all important.... The daily chores, the demands, the inspections, become the reality, not the voyage, nor the destination. (p. 340) END OF BLOCK QUOTE

In a similar vein, I have walked into many science classrooms and asked students what they were doing, only to be told something like "page 34," as if that were all I needed to know. For many students, school is just a series of tasks whose purposes are unclear and even what counts as success is mysterious, especially for students from less-advantaged backgrounds.

One study that is particularly notable for reducing the achievement gap between lower- and higher-achieving students was conducted by White and Frederiksen (1998). The study involved three teachers, each of whom taught 4 parallel seventh-grade classes in two U.S. schools. The average size of the classes was 31 students. In order to assess the representativeness of the sample, all the students in the study were given the Iowa Test of Basic Skills (ITBS), and their scores were close to the national average. All 12 classes followed a novel curriculum (called ThinkerTools) for 14 weeks. The curriculum had been designed to promote thinking in the science classroom through a focus on a series of seven scientific investigations (approximately 2 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. 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 one another feedback.

When the researchers analyzed the achievement of the students, the weakest students in the "reflective assessment" group performed as well as the best students in the control group, and the other students did even better, the result of which was to reduce by half the achievement gaps in the "reflective assessment" classes.

4. Activate students as owners of their own learning

One of the great traps of teaching is the belief that teachers create learning. This is particularly important when teachers are under pressure to improve student results, because studies have shown that when teachers are told they are responsible for making sure that their students do well, the quality of their teaching deteriorates, as does their students' learning (Deci et al. 1982); hence the old joke about schools being places where children go to watch teachers work.

Only learners create learning, and so, when we look at the role that assessment plays in promoting learning, the crucial feature is not the validity of the assessment, or its reliability, but its impact on the student. No matter how reliable or valid the assessment is, if it communicates to students that they cannot learn, it will hinder learning. Particularly important here is the work of Carol Dweck, who over a 30-year period has examined the way that students make sense of their successes and failures in school (see Dweck 2000, for a very readable summary of this huge volume of work). As a result of their experiences, some students come to believe that ability is fixed. The reason that this is so injurious to future learning is that every time students with this belief are faced with a challenging task, their first reaction is to engage in a calculation about whether they are likely to succeed or not. If they feel confident that they will succeed, or if they feel confident that the task is so hard that many others will also fail, they will attempt the task. However, if they feel that there is a danger that they will fail while others will succeed, they will disengage in order to protect their sense of self. Put simply, they are deciding that they would rather be thought lazy than dumb. It is the same choice that most adults would make.

There are other students, who, for a variety of reasons, have come to regard ability as incremental rather than fixed. They believe that "smart" is not something you *are* but something you *get*. For these students, challenging tasks are opportunities to increase their abilities, so whether their beliefs in their chances of success are high or low, they engage with a task in order to grow. What is particularly interesting is that the same student can believe the ability in science is fixed, while seeing ability in sports or music as incremental. Most students believe that ability in, for example, triple jump, throwing

the javelin, or guitar playing can be improved by practice. We need to inculcate the same beliefs about science.

In general, we need to activate students as owners of their own learning, so that they see challenge as a spur to personal growth, rather than as a threat to self-image. We need students who own their learning to the extent that they can self-manage both their emotional and their cognitive responses to challenge, so that all their energies are spent on developing capability rather than disguising its absence (see Wiliam 2007 for a summary of research in this area).

5. Activate students as learning resources for one another

The research on collaborative learning is one of the success stories of education research. Research in many areas of education produces ambiguous or contradictory findings whereas the research on collaborative learning, most notably the work of Robert Slavin (Slavin, Hurley, and Chamberlain 2003), has shown that activating students as learning resources for one another produces some of the largest gains seen in any educational interventions, provided two conditions are met. The first is that the learning environment must provide for group goals, so that students are working *as* a group, rather than just working *in* a group. The second is individual accountability, so that each student is responsible for his or her contribution to the whole, so there can be no "passengers."

With regard to assessment, then, a crucial feature is that the assessment encourages collaboration among students while they are learning. To achieve this, the learning intentions and success criteria must be accessible to the students (see above) and the teacher must support the students as they learn how to help each other improve their work. One particularly successful format for doing this has been the idea of "two stars and a wish." The idea is that when students are commenting on each others' work, they do not give evaluative feedback, but instead have to identify two positive features of the work (two "stars") and one feature that they feel merits further attention (the "wish").

Teachers who have used this technique with students as young as five years old have been astonished to see how appropriate the comments are, and, because the feedback comes from a peer rather than someone in authority over them, the recipient of the feedback appears to be more able to accept the feedback (in other words, students focus on growth rather than preserving well-being). In fact, teachers have told us that the feedback that students give each other, while accurate, is far more hard-hitting and direct than they would themselves feel able to provide. Furthermore, the research shows that the person providing the feedback benefits just as much as the recipient, because he or she is forced to internalize the learning intentions and success criteria in the context of someone else's work, which is less emotionally charged than doing so in the context of one's own work.

[A-Head] The "Big Idea": Keeping Learning on Track

The "big idea" that ties these strategies together is that assessment should be used to provide information to be used by students and teachers that is used to modify instruction

in real time in order to better meet student needs. In other words, assessment is used to keep learning on track.

That this is not common practice can be seen by imagining what would happen if an airline pilot navigated the way that most teachers teach. The pilot would set a course from the starting point (say New York) to the destination (say San Francisco). The pilot would then fly on this heading for the calculated time of travel, and then, when that time had elapsed, would land the plane at the nearest airport, and upon landing ask, "Is this San Francisco?" Worse, even if the plane had actually landed in Sacramento, the pilot would require all the passengers to leave, because he had to get on to his next job.

This would be absurd, and yet, this is how many teachers teach. They teach a unit for two or three weeks, and at the end of that teaching, they assess their students. And whatever the result of that assessment, the teacher is then on to the next unit, because of the district's pacing guide. If we are to keep learning on track, assessment cannot wait until the end of the unit. Instead, like the pilot, the teacher plans a course but then takes frequent readings along the way, adjusting the course as conditions dictate.

Substantial increases in student achievement are possible if we can increase the amount of assessment for learning in classrooms, but achieving this is no easy task.

[A-Head] Putting It Into Practice: The Case for Teacher Learning Communities

The fact that we know what needs to be done does not, of course, mean that we know how to do it. While the work of Black et al. (2003) has shown what we can achieve, the track record of professional development in producing significant effects on a large scale is rather unimpressive. However, this should not worry us unduly because very little of the professional development that teachers have received in the past is consistent with what we know makes for effective teacher change (Wiliam and Thompson 2007).

Why this is the case is complex, and beyond the scope of this article (see Wiliam 2003 for an extended discussion). What is clear is that in general, researchers have underestimated the complexity of what it is that teachers do; in particular, researchers have failed to understand how great an impact context has on teachers' practices. That is why "What works?" is not the right question, because everything works somewhere, and nothing works everywhere. The right question is "Under what conditions will this work?" And even if we might be able to answer this question scientifically at some point in the future, we are so far away from an answer now that we have to rely on the professional judgment of teachers.

This is why we cannot tell teachers what to do. This conclusion does not stem from a desire to be nice to teachers. Indeed, if I could identify a way of telling teachers what to do that would raise student achievement, I would have no hesitation in mandating it if I had the power to do so. Schools exist for the students, not to provide employment opportunities for teachers. The reason that we cannot tell teachers what to do is that we cannot provide them with reliable guides to action. The situations that they face in their classrooms are just too varied for us to predict. That is why professional judgment is

important; we have to develop the ability of teachers to react appropriately to situations for which they have not been specifically prepared.

The specific changes that I am arguing for here appear to be quite difficult for teachers to implement because they involving changing habits. A teacher with 20 years experience may well have asked something like half a million questions in her career. And when you've done something the same way half a million times, it's quite difficult to start doing it in any other way. But there is a deeper reason why change is difficult, even for inexperienced teachers. Teachers learn most of what they know about teaching before their 18th birthday. In the same way that most of us learn what we know about parenting through being parented, teachers internalize the "scripts" of school as students. Even the best four-year teacher education programs will find it hard to overcome the models of practice that student teachers will have learned in the 13 or 14 years that they spent in school.

This is why, if we are to have any chance of really changing teacher practice, we have to take seriously how hard this is going to be. We are asking teachers to change the routines and the practices that get them through the day; in the transition, they will get worse before they get better. Indeed, many of the teachers we have worked with described making these changes as "scary." They saw involving the students more in their learning as requiring them to "give up control" of their classrooms. A year later, the same teachers described the process as one of sharing responsibility for learning with the learners—the same process, but viewed in a radically different perspective.

The process that I believe provides the best mechanism for supporting teachers in making these changes is through the use of teacher learning communities (TLCs). I say this not out of some ideological commitment to the benefits of teachers talking to each other, but because of the nature of the changes we are seeking to produce. If we were trying to increase teacher subject knowledge, then TLCs would not be a very sensible approach—it would be far better to arrange for high-quality direct instruction. But when we are trying to change deeply ingrained, routinized practices or habits, then it seems that TLCs offer the best hope, and indeed, the results we have achieved in the United States have been very encouraging.

Over the last three years, we have tried a number of different approaches to establishing and sustaining TLCs, and as a result of this experimentation, it appears to us that five principles appear to be particularly important: gradualism, flexibility, choice, accountability, and support.

[B-Head] *Gradualism*

Asking teachers to make wholesale changes in their practices is a little like asking a golfer to change her swing during a tournament. Teachers have to maintain the fluency of their classroom routines, while at the same time disrupting them. Teachers should develop an action plan that specifies a small number of changes—ideally two or three—that they will make in their teaching. As teachers establish new techniques in their practices, they can take on additional ones. For administrators, there will be a temptation

to push teachers to change faster than they might otherwise do, but the result will only be a shallow adoption of the new practices as long as the teachers are being monitored. As soon as the supervision is relaxed, the teachers will revert to their earlier practices, and nothing will have been achieved. Even districts in a hurry will have to hasten slowly.

[B-Head] *Flexibility*

Teachers will need to modify techniques to make them work in their classrooms; in the process of adapting techniques, teachers often refine and improve them. One high school mathematics teacher heard about the "traffic light" technique in which at the end of a piece of work, students indicate their confidence in their understanding of a piece of work with a green, yellow, or red circle, representing complete, partial, or little understanding. She decided that she would not wait until the end of the lesson to engage students in this self-assessment, and gave each student a disc, green on one side and red on the other. At the start of each lesson students place the disk on their desk with the green side showing. A student can indicate confusion at any time by turning the disc over to show red. The teacher found that students who had never asked a question all year in class were prepared to signal their confusion in this way.

Another teacher tried this approach, but found it difficult to see the discs from the front of the class so she provided each student with three paper cups, one green, one yellow, and one red, nested inside each other on the students' desks. Students used these cups to indicate whether they were following the teacher's explanation (green), wanted the teacher to slow down (yellow), or wanted her to stop in order to ask a question (red). The teacher made students accountable for signaling correctly by establishing a rule that whenever one student showed red, a student who was showing green or yellow would be chosen at random to come to the front of the class to answer the question posed by the student showing red. In this classroom there is nowhere to hide!

{B-Head] *Choice*

As noted above, teachers often describe the process of changing their practices as "scary," but when they are responsible for choosing what they will change about their practices they feel empowered, especially when they can choose from a range of techniques those that appeal to them. This choice lies, however, within a framework of accountability. While teachers are free to choose what they change, they are accountable for changing something.

[B-Head] *Accountability*

Most professionals involved in teacher development will have had the experience of generating considerable enthusiasm for, and commitment to, change during a workshop, only to find that all the good intentions seem to be erased once the teachers return to the classroom. Teachers should be held accountable for making changes by colleagues at monthly meetings of their teacher learning community. Each teacher describes what he or she tried and how it went. Teachers repeatedly tell us that having to face their colleagues helps them move their "change" task to the top of their in-box.

[B-Head] *Support*

Along with ideas for what to change and the support of a teacher learning community, two elements are highly desirable, if not essential, for teacher learning. The first is training for those who will lead the learning communities. The person leading the learning community must be clear about his or her role. The role of the leader is not to create teacher change but to engineer situations in which the teacher change can take place. Those in supervisory roles often find this more difficult than do teachers, because those teaching every day understand how difficult it is to change practice.

The second element is peer observation. Collaborative planning in the monthly TLC meetings can help teachers focus on what they want to develop in their practices, but teachers need support in carrying out these resolutions. To distinguish these observations clearly from those routinely carried out to manage performance, these observations should be done by peers rather than supervisors. The teacher being observed must set the agenda for the observation and spell out for the observer what should count as evidence. By defining the observer's role, both in terms of what is to be looked for and what counts as evidence, the observer's own prejudices are minimized, and the difference between this and supervisory observation is emphasized. For further details on setting up, and sustaining teacher learning communities for formative assessment, see Wiliam (2007/2008).

[A-Head] Integrated Assessment

The available research evidence, as well as our experience of working with many school districts in the United States, suggests that the use of teacher learning communities, focused on the use of minute-to-minute and day-to-day assessment to adjust instruction to meet student needs, represents the most powerful single approach to improving student achievement. However, if we are to maximize the impact on student learning, other parts of the system need to be "in sync." As well as the minute-by-minute and day-by-day assessments that allow teachers to keep learning on track, teachers also need a range of more formal assessment tasks and activities that support valid and reliable conclusions about the extent of student learning. Our experience is that teacher-made assessments often focus on shallow aspects of learning, rather than the 'big ideas'. Developing high-quality assessments that involve students and motivate them to improve takes time, but there are good examples of how to go about this (Stiggins et al. 2004). On a longer timescale, quarterly assessments that are paced to the curriculum can provide school leadership teams with valuable information about the progress—or lack of it—that is being made by students, and annual diagnostic analyses of high-stakes tests can provide important insights into the alignment between the teaching and the national curriculum. None of these different kinds of assessment is in conflict with any of the others. Each represents an important part of complex machine providing information at the right level of specificity for the decision that needs to be made. Together they form a balanced assessment system that can produce unprecedented increases in science achievement, benefiting both the individual and society as whole.

References

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

Clymer, J. B., and D. Wiliam. 2006/2007. What's wrong with the way we grade science? *Educational Leadership* 64(4): 36–42.

Deci, E. L., N. H. Speigel, R. M. Ryan, R. Koestner, and M. Kauffman. 1982. The effects of performance standards on teaching styles: The behavior of controlling teachers. *Journal of Educational Psychology* 74: 852–859.

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

Hayes, V. P. 2003. *Using pupil self-evaluation within the formative assessment paradigm as a pedagogical tool.* EdD thesis, University of London.

Hill, H. C., B. Rowan, and D. L. Ball. 2005. Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal* 42(2): 317–406.

Hoff, D. J. 2007. Economists tout value of reducing dropouts. *Education Week* 26(Feb. 14): 5, 15.

Jenkins, A., R. Levacic, and A. Vignoles.. 2006. *Estimating the relationship between school resources and pupil attainment at GCSE* (Vol. RR727). London, UK: Department for Education and Skills.

Jepsen, C., and S. G. Rivkin. 2002. *What is the tradeoff between smaller classes and teacher quality?* Cambridge, MA: National Bureau of Economic Research.

Kluger, A. N., and A. DeNisi. 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.

Lleras-Muney, A. 2005. The relationship between education and adult mortality in the United States. *Review of Economic Studies* 72(1): 189–221.

Rodriguez, M. C. 2004. The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education* 17(1): 1–24.

Slavin, R. E., E. A. Hurley, and A. M. Chamberlain. 2003. Cooperative learning and achievement. In W. M. Reynolds and G. J. Miller (Eds.), *Handbook of psychology, volume 7: educational psychology* (pp. 177–198). Hoboken, NJ: Wiley.

Stiggins, R. J., J. A. Arter, and S. Chappuis. 2004. *Classroom assessment for student learning: Doing it right—using it well*. Portland, OR: Assessment Training Institute.

White, M. A. 1971. The view from the student's desk. In M. L. Silberman (Ed.), *The experience of schooling* (pp. 337–345). New York: Rinehart and Winston.

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

Wiliam, D. 2003. The impact of educational research on mathematics education. In A. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, and F. K. S. Leung (Eds.), *Second international handbook of mathematics education* (pp. 469–488). Dordrecht, Netherlands: Kluwer Academic Publishers.

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

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

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

Wilson, M., and K. Draney. 2004. Some links between large-scale and classroom assessments: The case of the BEAR assessment system. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education* (Part II, pp. 132–154). Chicago, IL: University of Chicago Press.

Further Reading

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

Black, P., and C. Harrison. 2002. *Science inside the black box: Assessment for learning in the science classroom*. London, UK: NFER-Nelson.

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

Clymer, J. B., and D. Wiliam. 2006. Improving the way we grade science. *Educational Leadership* 64(4): 36–42.

Leahy, S., C. Lyon, M. Thompson, and D. Wiliam. 2005. Classroom assessment: Minute-by-minute and day-by-day. *Educational Leadership* 63(3): 18–24.

Wiliam, D. 2006. Assessment: Learning communities can use it to engineer a bridge connecting teaching and learning. *Journal of Staff Development* 27(1): 16–20.

Wiliam, D. 2006. Assessment for learning: Why, what and how. *Orbit: OISE/UT's magazine for schools* 36(3): 2–6.

Wiliam, D. (2007/2008, December/January). Changing classroom practice. *Educational Leadership,* **65**(4)*,* 36-42.