From teachers to schools: scaling up professional development for formative assessment

Siobhan Leahy (Edmonton County School, Enfield, UK) &
Dylan Wiliam (Institute of Education, University of London)

Introduction

The surgeon Atul Gawande (2007) has pointed out that the advances in highly demanding aspects such as surgery have been greater than in apparently “easier” areas such as basic hygiene. He quotes the example of washing hands after contact with patients, where the compliance rate in many hospitals is below 50% even though it is widely accepted that compliance rate of over 95% are required to control the spread of resistant forms of staphylococcus aureus such as vancomycin-resistant Staphylococcus aureus (VRSA) and methicillinin-resistant Staphylococcus aureus (MRSA) (p. 15).

The problem is not ignorance, nor willful disobedience. A strict procedure is specified for hand washing, but Gawande points out,

> Almost no one adheres to this procedure. It seems impossible. On morning rounds, our residents check in on twenty patients in an hour. The nurses in our intensive care units typically have a similar number of contacts with patients requiring hand washing in between. Even if you get the whole cleansing process down to a minute per patient, that’s still a third of staff time spent just washing hands. Such frequent hand washing can also irritate the skin, which can produce dermatitis, which itself increases bacterial counts. (p. 18)

An analysis of the practice of teachers would appear to share many features of that of clinicians. The work moves at a fast pace, so that there is little time for reflective thought, and as a result, the workplace behaviors are driven by habit as much as anything else. This would not matter too much if those habits were the most effective ways of engendering student learning, but there is ample evidence that a significant proportion of teacher practices are sub-optimal, and that significant improvements in student achievement would be possible with changes in teachers’ classroom practices. In this paper, we describe our efforts to develop ways of supporting changes in teachers’ classroom practice through a focus on formative assessment, with a particular focus on finding ways of doing so that could, in principle, be implemented at scale—for example across 300,000 classrooms in England, or across 2 million classrooms in the United States. In the following two sections, we briefly summarize the research that has led us to focus on formative assessment as the most powerful lever for moving teacher practice in ways that are likely to benefit students, and why we have adopted teacher learning communities as the mechanism for supporting teachers in making these changes in their practice. In subsequent sections we describe the creation of a professional development product that would assist the creation of school-based teacher learning communities to support teachers in taking forward their own formative assessment practices, and we conclude with a case study of the implementation of the product in one urban school district in London.

The case for formative assessment

The evidence that formative assessment is a powerful lever for improving outcomes for learners has been steadily accumulating over the last quarter of a century. Over that time, at least 15 substantial reviews of research, synthesizing several thousand research studies, have documented the impact of classroom assessment practices on students (Fuchs & Fuchs 1986; Natriello, 1987; Crooks, 1988; Bangert-Drowns, Kulik, Kulik & Morgan, 1991; Dempster, 1991, 1992; Elshout-Mohr, 1994; Kluger & DeNisi, 1996; Black & Wiliam, 1998; Nyquist, 2003; Brookhart, 2004; Allal & Lopez, 2005; Köller, 2005; Brookhart, 2007; Wiliam, 2007; Hattie & Timperley, 2007; Shute, 2008).

While many of these reviews have documented the negative effects of some assessment practices, they also show that, used appropriately, assessment has considerable potential for enhancing student achievement. Drawing on the early work of Scriven (1967) and Bloom (1969), it has become common to describe the use of assessment to improve student learning as "formative assessment" although more recently the phrase "assessment for learning" has also become common. In the United States, the term "assessment for learning" is often mistakenly attributed to Rick Stiggins (2002), although Stiggins himself has always attributed the term to authors in the United Kingdom. In fact, the earliest use of this term in this sense appears to be a paper given at the annual conference of the Association for Supervision and Curriculum Development (James, 1992) while three years later, the phrase was used as the title of a book (Sutton, 1995). However, the first use of the term "assessment for learning" in contrast to the term "assessment of learning" appears to be Gipps & Stobart (1997), where these two terms are the titles of the second and first chapters respectively. The distinction was brought to a wider audience by the Assessment Reform Group in 1999 in a guide for policymakers (Broadfoot, Daugherty, Gardner, Gipps, Harlen, James & Stobart, 1999).

Wiliam (2009) summarizes some of the definitions for formative assessment (and assessment for learning) that have been proposed over the years, and suggests that the most comprehensive definition is that adopted by Black and Wiliam (2009):

> Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited. (p. 9)

In commenting on this definition, Black and Wiliam emphasize that the focus of the definition on *decisions* represents a compromise between basing the definition on intent (which would be a loose definition, admitting almost any data collection activity as formative) and basing it on outcomes (which would be a highly restrictive definition, due to the unpredictable nature of learning). The second point they make is that while these decisions are regularly made by the teacher, it is also the case that the learners themselves, and their peers, may also be involved in making these decisions. Indeed, ultimately, as the work of Monique Boekaerts suggests, unless the learners themselves choose growth in learning over personal well-being, little learning is likely to take place (Boekaerts, 2006). The third point is that the definition is probabilistic, again due to the unpredictable nature of learning, and the fourth point is that formative assessment need not alter instruction to be formative—it may simply confirm that the proposed course of action is indeed the most appropriate.

The general finding is that across a range of different school subjects, in different countries, and for learners of different ages, the use of formative assessment appears to be associated with considerable improvements in the rate of learning. Estimating how big these gains might be is difficult because most of the reviews appear to ignore the fact that outcome measures differ in their sensitivity to instruction (Wiliam, 2009) but it seems reasonable to conclude that use of formative assessment can increase the rate of student learning by somewhere between 50 and 100 per cent. This suggests that formative assessment is likely to be one of the most effective ways—and perhaps the most effective way—of increasing student achievement (Wiliam & Thompson, 2007, for example, estimate that it would be 20 times more cost-effective than typical class-size reduction programs).

The substantial evidence regarding the potential cost-effectiveness of formative assessment as a lever for school improvement has, predictably, attracted considerable attention, and a number of test publishers have produced what they call "formative assessment systems" or "benchmark assessments." These include the *MAP* produced by the Northwest Evaluation Association (NWEA), the *Focus on Standards*™/*Instructional Data Management System*™ produced by ETS, *Homeroom*™ produced by Princeton Review, *Benchmark Tracker*™/*SkillWriter*™ and *Stanford Learning First*™ by Harcourt Assessment, and *Prosper*™ produced by Pearson Assessments, as well as a host of other similar systems. Typically, these systems provide for assessment of student progress at regular intervals (generally every four to nine weeks) and provide reports that identify students, or particular aspects of the curriculum, that require special attention.

While some of the publishers of these products simply appropriate the existing literature on formative assessment as evidence of their efficacy, others have undertaken original research on the impact of these formative assessment systems. ETS, as the owner of "Focus on Standards™," has undertaken investigations of the impact of adoption of this program and found that while it could have significant impact in particular settings (for example when the alignment of curriculum to standards was poor) the general impact appears to be limited (Goe & Bridgeman, 2006).

In terms of the definition proposed by Black and Wiliam discussed above, such systems may be formative, in the sense that they may provide evidence about student achievement that could be used to make better decisions about instruction than would have been possible without that evidence. However, since few if any of the studies synthesized in the 15 reviews mentioned earlier dealt with such "long-cycle" formative assessment, one cannot conclude on the basis of the existing research that these periodic assessments are likely to have significant impact on student achievement. Formal systems for testing students on a regular basis may have a role to play in the effective monitoring of student progress—indeed, some means of tracking student progress over the medium term, and taking action to address any problems identified—would seem to be an essential component of any comprehensive assessment system. But it is disingenuous at least, and possibly mendacious, to claim that the research literature provides evidence of the effectiveness of such systems. Quite simply, it does not (Popham, 2006; Shepard, 2007). That is not to say that such evidence will not be forthcoming in the future—it may well be—but little such evidence has been assembled to date.

The same can be said for what are called "common formative assessments" or "interim assessments," defined by DuFour, DuFour, Eaker & Many (2005) as:

> An assessment typically created collaboratively by a team of teachers responsible for the same grade level or course. Common formative assessments are frequently administered throughout the year to identify (1) individual students who need additional time and support for learning, (2) the teaching strategies most effective in helping students acquire the intended knowledge and skills, (3) program concerns—areas in which students generally are having difficulty achieving the intended standard— and (4) improvement goals for individual teachers and the team. (p. 214)

Again, while such assessments clearly have a valuable role to play in aligning instruction to standards, as a focus for professional dialogue, and for supporting good management and supervision, the evidence on the impact of such "medium-cycle" formative assessments on student achievement is weak.

In contrast, there is strong evidence that what Wiliam and Thompson (2007) term "short-cycle" formative assessments—can have a profound impact on student achievement. Yeh (2006) summarizes a number of studies that show that what he calls "rapid formative assessment" (assessments conducted from two to five times per week) can significantly improve student learning. On an even shorter time-scale Black, Harrison, Lee, Marshall and Wiliam (2003) describe how they supported a group of 24 mathematics and science teachers in developing their use of "in-the-moment" formative assessment and found that even when measured through externally-set, externally-scored, state-mandated standardized assessments, the gains in student achievement were substantial, equivalent to an increase of the rate of student learning of around 70% (Wiliam, Lee, Harrison & Black, 2004). Other similar interventions have produced similar effects (Hayes, 2003; Clymer & Wiliam, 2006/2007).

It therefore seems reasonably clear that the effects that the literature shows are possible are indeed achievable in real classrooms, even where outcomes are measured using externally-mandated, standardized tests. What is much less clear is how to achieve these effects at scale—across 300,000 classrooms in England, or across 2 million classrooms in the United States.

Designing for scale

In designing ways of supporting the implementation of formative assessment across a large number of classrooms, we and our colleagues at the Educational Testing Service adopted as a design constraint the idea of "in-principle scalability." By this we meant that the intervention need not be scalable at the outset, but any aspect of the intervention that could not, under any reasonable set of assumptions, be implemented at scale, was ruled out.

A second constraint was a commitment to a single model for the whole school. One of the most surprising findings in our work with schools over the past 20 or so years is how 'Balkanized' the arrangements for teacher professional development are, especially in secondary schools. It is quite common to find the mathematics teachers engaged in one set of professional development activities, the science teachers another, and the social studies teachers doing something else entirely. Quite apart from the fact that this is difficult and confusing for the students, these differences in approach make it far more difficult to generate a coherent dialogue around the school about teaching and learning. However, while we were committed to a single model for the whole school, we realized we had also to honor the specificities of age and school-subject. Teaching five-year-olds is not the same as teaching ten-year-olds, and teaching mathematics is not the same as teaching history.

We were also aware that any model of effective, scalable teacher professional development would need to pull off a delicate balancing act between two conflicting requirements. The first was the need to ensure that the model was sufficiently flexible to allow the model to be adapted to the local circumstances of the intervention, not just to allow it to succeed, but also so that it could capitalize upon any affordances present in the local context that would enhance the intervention. The second was to ensure that the model was sufficiently rigid to ensure that any modifications that did take place preserved sufficient fidelity to the original design to provide a reasonable assurance that the intervention would not undergo a "lethal mutation" (Haertel, cited in Brown & Campione, 1996).

To address this issue, we explicitly adopted a framework entitled "tight but loose":

> The Tight but Loose formulation combines an obsessive adherence to central design principles (the "tight" part) with accommodations to the needs, resources, constraints, and particularities that occur in any school or district (the "loose" part), *but only where these do not conflict with the theory of action of the intervention.* (Thompson & Wiliam, 2008 p.35; emphasis in original).

A fuller account of the application of the "Tight but Loose" framework to the design of a professional development program for supporting teachers in their use of formative assessment can be found in Thompson and Wiliam (2008). Of particular relevance here is that our design work was guided by a number of principles which previous work had suggested were important in supporting teachers in the development of their practice of formative assessment in particular: choice, flexibility, small steps, accountability and support (Wiliam, 2006).

Choice

It is often assumed that to improve, teachers should work to develop the weakest aspects of their practice, and for some teachers, these aspects may indeed be so weak that they should be the priority for professional development. But for most teachers, our experience has been that the greatest benefits to students come from teachers becoming even more expert in their strengths. In early work on formative assessment with teachers in England (Black, Harrison, Lee, Marshall & Wiliam, 2003), one of the teachers, Derek (this, like the names of all teachers, schools, and districts mentioned in this paper, is a pseudonym) was already quite skilled at conducting whole-class discussion sessions, but he was interested in improving this practice further. A colleague of his at the same school, Philip, has a much more “low-key” presence in the classroom, and was much more interested in helping students develop skills of self-assessment and peer-assessment. Both Derek and Philip are now extraordinarily skilled practitioners—amongst the best we have seen—but to make Philip work on questioning, or to make Derek work on peer-assessment and self-assessment would, we feel, be unlikely to benefit their students as much as supporting each teacher to become excellent in their own way. Furthermore, we have found that when teachers themselves make the decision about what it is that they wish to prioritize for their own professional development, they are more likely to “make it work”. In traditional ‘top-down’ models of teacher professional development, teachers are given ideas to try out in their own classrooms, but often respond by blaming the professional developer for the failure of new methods in the classroom (e.g., “I tried what you told me to do and it didn’t work”). However, when the choice about the aspects of practice to develop is made by the teacher, then the responsibility for ensuring effective implementation is shared.

Flexibility

Teachers need the flexibility to be able to modify or “morph” the formative assessment techniques with which they are presented to fit their own classroom context (Ginsburg, 2001). The danger in this is that a teacher may so modify an idea that it is no longer effective (an example of the “lethal mutation” described above). What is needed, therefore, is a way of allowing teachers flexibility, while at the same time constraining the use of that flexibility so that modifications to the original ideas do not unduly weaken their effectiveness. Our solution to this was a clarification of the distinction between the *strategies* and the *techniques* of formative assessment. To define the strategies of formative assessment, we began by identifying three key processes involved in formative assessment:

Identifying where the learners are in their learning
Identifying where they are going
Identifying what steps need to be taken to get there.

Considering the role of the teacher, the learners, and their peers in these processes yielded nine cells, which can be collapsed into the five “key strategies” of formative assessment as shown in Figure 1 (Wiliam & Thompson, 2007).

Each of these five “key strategies” provides a focal point for a range of related aspects of practice for teachers. In other words, they provide a starting point for a consideration of a number of wider issues in teaching, such as curriculum, psychology, and pedagogy.

For each of these five key strategies, a number of specific classroom routines, termed “techniques,” were identified. The techniques take the form of “validated practices” that are consistent with the research on formative assessment and teachers are free to adopt whichever of these techniques they wish (thus providing choice—see above). By anchoring the techniques to (at least) one of the five key strategies, we provided a means by which teachers could modify the techniques, but still provide a reasonable assurance of fidelity to the original research.

Figure 1. Aspects of formative assessment

	Where the learner is going	Where the learner is right now	How to get there
Teacher	Clarifying learning intentions and sharing and criteria for success (1)	Engineering effective classroom discussions, activities and tasks that elicit evidence of learning (2)	Providing feedback that moves learners forward (3)
Peer	Understanding and sharing learning intentions and criteria for success (1)	Activating students as instructional resources for one another (4)	
Learner	Understanding learning intentions and criteria for success (1)	Activating students as the owners of their own learning (5)	

Small steps

In implementing any professional development model, we have to accept that teacher learning is slow. In particular, for changes in practice as opposed to knowledge to be lasting, they must be integrated into a teacher's existing routines, and this takes time. Many of those involved in professional development are familiar with the experience of encouraging teachers to try out new ideas, and seeing them being enacted when they visit teachers' classrooms, only to learn shortly afterwards that the teachers have reverted to their former practices.

Some authors have attributed such reversion to resistance on the part of teachers, caused by each teacher's adherence to a series of professional habits, that to a very real extent, represent a core part of each teacher's professional identity. This may be part of the reason for the slowness of teacher change, but it seems to us that a far more significant cause of the failure of many changes in classroom practice to "stick" is due to the fact that high-level performance in a domain as complex as teaching requires automatizing a large proportion of the things that teachers do. For learner drivers, shifting gear, using the turn indicator and steering all at the same time seem impossibly complicated—and undertaken consciously, they are. Experienced drivers have practiced these activities so many times that they become automated and thus take up little of the available resources for cognitive processing. However as anyone who tries to change the way they drive—for example in order to reduce the extent to which they "ride the clutch"—has discovered, these automated procedures are extremely hard to change.

Teaching is even more extreme than driving in this respect, because every teacher comes to the profession with a series of "scripts" of how classrooms should operate "hard-wired" into their minds from their time as a student. These scripts, such as requiring students to raise their hands if they have an answer to a teacher's question, seem natural, but of course they are learned, and maladaptive (Wiliam, 2005).

Moreover, many of the changes in practice associated with implementing formative assessment are not just difficult to change because they are habituated—they also contradict widely-distributed and strongly-held beliefs about, for example, the value of grades for motivating students. Even when teachers are convinced of the value of approaches such as "comment-only grading" they are often dissuaded from trying them out by more senior colleagues who dismiss innovations as fads advocated by ivory tower academics who don't know what real teaching is. That is why, even if we are in a hurry to help teachers improve their practice, we should "hasten slowly".

Support and accountability

The last two principles—support and accountability—can be thought of as two sides of the same coin. Indeed, elsewhere, we have described them as a single feature of effective learning environments for teachers: supportive accountability (Ciofalo & Leahy, 2006). The central idea is the creation of structures that, while making teachers accountable for developing their practice, also provide the support for them to do this.

Clearly, creating this 'supportive accountability' could be done in a number of ways. One way would be to assign each teacher a coach, but this would be expensive, and it is by no means clear that an adequate supply of coaches would be available. The requirements of "in -principle scalability" led to the rejection of a coaching-based model, and instead, focused on the idea of building-based teacher learning communities.

Supporting formative assessment with teacher learning communities

Between 2003 and 2006, working with other colleagues at Educational Testing Service, we developed and piloted a number of models for supporting teachers (for extended accounts of these early developments, see Thompson & Goe, 2008; Wylie, Lyon & Goe, 2009; Wylie, Lyon, & Mavronikolas, 2008). One of our earliest models involved one of us (SL) meeting every two or three weeks with groups of four to six high school mathematics teachers to discuss the changes they were attempting to make in their practice. As a result of this work, it became clear that a two-week cycle did not allow enough time for the teachers involved to plan and implement changes in their practice in time for reporting back at the next meeting. In contrast, implementations that involved meetings that occurred at intervals of six weeks or more appeared to lose momentum. This led to the adoption of a monthly cycle of meetings that has persisted in all our implementations to the present day. In work with literally hundreds of schools in dozens of districts over the last five years, we have not come across any evidence that suggests that intervals between meetings of approximately four weeks is not an optimum, at least in respect of changes in practice related to formative assessment.

Originally we had assumed that schools would be able to find two hours for each of the monthly meetings, and while this was clearly possible in some districts, in others it was not, so we looked carefully at ways of reducing the length of the monthly meeting. After experimentation with different lengths of meetings (including meetings as short as 60 minutes), we concluded that 75 minutes should be an absolute minimum.

Our experiences with meetings with small numbers of participants had also led us to conclude that the minimum number of participants needed for a meeting to be effective was around eight. Meetings with fewer than eight participants often required significant input from the group's leader or facilitator, particularly when occasional absences due to illness and other factors reduced the size of the group further. While such intensive support from the facilitator might provide an effective learning environment for those attending, such a model would be unlikely to be scalable.

On the other hand, where the group was much larger than twelve (as was often the case in our early work in the Cleveland Municipal School District), there was not enough time to hear back from each member of the group. In interviews, many participants in teacher learning communities have told us that it was the fact that they knew that they would be required to give their colleagues an account of what they had been doing that made them prioritize working on changing their classroom practice over all the pressing concerns of a teacher's life (Ciofalo & Leahy, 2006).

As well as design guidelines for the size of group and frequency of meetings, we also explored the extent to which it was necessary for teachers to share particular assignments (e.g., early grades or subject-specialisms in secondary schools). It has been our experience that teachers greatly value meeting in mixed-subject groups, in order to get ideas from teachers of different subjects or different ages. However, we had also observed many instances of a teacher rejecting suggestions from other members of the group with a claim that the suggested idea would not work for her or his own subject-specialism. In order to balance these tensions, we have explored models where the teachers do not all come from the same subject specialism, but, in order to provide some disciplinary support, we ensure that for each teacher in the group, there was at least one other with the same age- or subject-specialism. To date, we do not have any data that suggests that any particular method of constituting a group is better than another, although we are aware that the scope for deep conversations about subject matter are likely to be limited where the group is made up of individuals with different subject specialisms (Grossman, Wineburg & Woolworth, 2000).

One final design feature of the monthly meetings of the teacher learning communities was related to their structure. As a result of Shulman's work on the "signature pedagogies" of the professions (Shulman, 2004), we realized that there could be considerable benefits of adopting a standard structure for these monthly meetings, but that, in most approaches to teacher professional development, novelty was often regarded as paramount, in order to keep things "fresh." In such situations, the result is often that the structure of the learning takes precedence over the content of the learning, and much time is spent on learning about the structure of the learning, rather than the learning itself. In order to "background" the structure of the learning, so that the learning itself could be foregrounded, we decided to arrange each meeting around the same six activities, that would occur in the same sequence in each monthly meeting. The fact that each meeting follows the same structure means that participants come to the meeting knowing the roles they are to play, both in terms of reporting back on their own experiences, and providing support to others.

Introduction (5 minutes)

Agendas for the meeting are circulated and the learning intentions for the meeting are presented.

Starter activity (5 minutes)

Participants engage is an activity to help them focus on their own learning.

Feedback (25 minutes)

Each teacher gives a brief report on what they committed to try out during the “personal action planning” section at the previous meeting, while the rest of the group listen appreciatively and then offer support to the individual in taking their plan forward.

New learning about formative assessment (20 minutes)

In order to provide an element of novelty into each meeting of the TLC, and to provide a steady stream of new ideas, each meeting includes an activity that introduces some new ideas about formative assessment. This might be a task, a video to watch and discuss, or a ‘book study’ in which teachers will discuss a book chapter relevant to formative assessment that they have read over the past month.

Personal action planning (15 minutes)

The penultimate activity of each session involves each of the participants planning in detail what they hope to accomplish before the next meeting. This may include trying out new ideas or it may simply be to consolidate techniques with which they have already experimented. This is also a good time for participants to plan any peer observations that they plan to undertake. It is our experience that if the participants leave the meeting without a definite date and time to observe one another, the peer observation is much less likely to take place (Maher & Wiliam, 2007).

Summary of learning (5 minutes)

In the last five minutes of the meeting, the group discusses whether they have achieved the learning intentions they set themselves. If they have not, there is time for the group to decide what to do about it.

From principles to products

The research on formative assessment and the design principles for the teacher learning summarized above create a very clear vision of what should be happening in classrooms, and what kinds of teacher professional development might help move teachers towards such practice. However, this clarity only takes one so far in the design of products that might be distributed at scale.

Since 2005, ETS has been involved in extensive piloting and trialing of a product called “Keeping Learning on Track™” and this will be launched later in 2009. Early versions of KLT involved a four-day training program (two days for all participants and two further days for TLC leaders), and the materials to support the training and the monthly TLC meetings ran to several hundred pages, providing detailed guidance to TLC leaders on which items on the agenda to omit if time was running short. While such extensive support materials might be necessary for some schools, there is a danger that such detail conveys the idea that the person leading the TLC needs to be an “expert” in formative assessment. One of the key drivers motivating the five principles identified above was that each teacher is a professional, making plans for her or his own personal professional development, and getting the support of like-minded professionals in carrying out those plans.

Therefore, since 2007, we (the two authors of the current paper) have been involved in a developing a “minimal” set of materials for supporting schools in supporting teachers in developing their own practice of formative assessment.

“Embedding formative assessment” (Leahy & Wiliam, 2009) is a CD-ROM containing materials for a school to run its own one-day workshop on formative assessment (PowerPoint slides and a complete transcript for the workshop), seven 10-minute video clips of one of us (DW) presenting the research basis of, and some techniques for, formative assessment, and materials for running nine monthly follow-up meetings of the TLC. The materials for supporting the nine monthly meetings run to 75 pages in total, including agendas, handouts, and notes for the group leader.

As the materials for the monthly meetings were developed, they were circulated to about 60 other schools that had expressed an interest in trying them out. Further materials were sent only to those schools that had provided feedback on the materials they had received.

Perhaps the most interesting finding from this early phase of development was that many schools appropriated elements of the program to support their existing plans. Despite having attended presentations where the research basis for formative assessment was discussed at some length, and where it was shown that there was little or no research to support other innovations in learning that were attracting interest at the time (e.g., “brain gym” or learning styles), many schools reported that they liked the idea of teacher learning communities, but had decided to use them to support teachers in whatever was the school’s current priority for professional development (e.g., differentiated instruction, personalization, learning styles). Of course, this appropriation of resources is hardly surprising, impossible to prevent, and may be very positive in its outcomes, but what was surprising was that most of those who had transformed the innovation beyond recognition appeared to believe that they were implementing the materials *as intended*.

Evaluation of both *Keeping Learning on Track*™ and *Embedding Formative Assessment* are both at early stages, but the challenges of implementing effective professional development at scale are manifest in both programs. In the remainder of this paper, we discuss briefly a case study of the implementation of *Embedding Formative Assessment* in a school district in England. While some of the details are unique to the English education system, the major problems appear to be exactly the same as are encountered in the United States.

A case study in one district

Cannington is a local authority (school district) in Greater London, covering an area of approximately 10 square miles, and serving a diverse population of approximately 200,000 residents, with three times as many of its residents from minority ethnic communities as the country as a whole.

In July 2007, a charitable foundation made available some funds for the establishment of a research project designed to raise student achievement in mathematics, science, and modern foreign languages—subjects that supervisory staff in Cannington had identified as priorities for development. Although the *Embedding formative assessment* materials had been intended for whole-school use, the opportunity to examine their use in subject-specific learning communities across an entire school district was considered to be too important to pass up.

In November 2007, a presentation was made to a meeting of the principals of the secondary schools in Cannington, proposing the establishment of three teacher learning communities in each secondary school, one focusing on mathematics, one focusing on science and the third, on modern foreign languages, to provide support for teachers in their development of classroom formative assessment practices. Members of the project team attended meetings of the Cannington principals over the subsequent months to provide updates on progress, and in July 2008, a series of three training events was held—one for each school subject—for teachers in the participating schools. The number of teachers from each school attending each of the events is shown in table 1.

	Number of teachers attending training event for:		
School	Mathematics	Science	Modern Languages
Ashtree School	1	1	0
Cedar Lodge School	5	1	3
Hawthorne School	4	10	5
Hazeltree School	7	12	2
Larchtree School	1	0	0
Mallow School	6	7	3
Poplar School	11	3	1
Spruce School	7	8	5
Willowtree School	2	5	2
Totals	44	47	21

Table 1: Numbers of teachers attending each training event

The training day consisted of a brief summary of the research on formative assessment, an overview of the five “key strategies” of formative assessment, an introduction to approximately 30 different techniques that teachers might use to implement formative assessment in their classrooms, and details on the creation of the three subject-specific school-based teacher learning communities that would be established to provide ongoing support in each school. The training session also provided guidance on the role of the leader of each of the three teacher learning communities to be established in each school.

The reactions of the teachers to the training was extremely positive, and at the end of the training day, the participants from six of the nine schools appeared to have a firm plan for implementation. One school (Ashtree) had decided to delay participation in the project for a year, and Hazeltree School had earlier that month decided to create mixed-subject, rather than subject-specific teacher learning communities, as they felt that this was more in keeping with the professional development work that had already taken place at the school.

However, since the funding had been provided specifically for supporting the teacher of mathematics, science and modern foreign languages, it was agreed that this would in effect mean that Hazeltree would be withdrawing from the project, although they continued to receive all materials necessary for supporting teacher learning communities. Larchtree School had only sent a single teacher (to the mathematics session), but the teacher concerned appeared confident that she would be able to "cascade" the training to other teachers in the mathematics department, and possibly also to the other subjects.

While it was not possible for each teacher of mathematics, science and modern foreign languages in each school to attend the one-day workshop, all teachers of these subjects in the secondary schools in Cannington were provided with a short (30-page) booklet outlining the major principles of formative assessment, together with specific applications in their subject (Hodgen & Wiliam, 2006; Black & Harrison, 2002; Jones & Wiliam, 2007).

In order to provide a simple channel of communication between the teachers in the school and the project, 6 expert facilitators (two for each subject-specialism) were appointed. Their major role was not to "drive" the implementation of the program in the schools, but rather to respond to requests from TLC leaders and administrators within the school on the use of the materials. Each facilitator kept a log of their contacts with teachers at the school, which provided the main source of evidence on the extent to which the TLCs were functioning as intended.

Given the involvement of the principals of the school at each step of the process up to this point, and their investment in releasing significant numbers of teachers to attend the initial workshops, we expected the teacher learning communities would be established quickly, but for a variety of reasons, adoption was extremely patchy.

At the end of seven months, logs provided by the facilitators were coded by two different raters, with each rater being asked to rate the progress made by each teacher learning community on a scale from 1 (little or no progress) to 4 (good progress). When the ratings generated independently were compared, in no case did the ratings differ by more than one point, and agreed ratings of two raters are shown in table 2.

	Progress		
School	Mathematics	Science	Modern Languages
Ashtree School*	-	-	-
Cedar Lodge School	1	1	2
Hawthorne School	2	2	4
Hazeltree School*	-	-	-
Larchtree School	4	1	1
Mallow School	3	1	2
Poplar School	1	3	3
Spruce School	4	3	3
Willowtree School	1	1	4
Average	2.3	1.8	2.8

*Did not participate in project (Ashtree deferred for a year, Hazeltree implemented a different model).

Table 2: Extent of progress of teacher learning communities in each school

Two features of this table are particularly worth noting. First the greatest progress appears to have been made in Modern Foreign Languages, which (we hope coincidentally!) is the subject with the least-well attended initial session. Second, with the exception of Spruce School, there does not appear to be any tendency for the best-progressing TLCs to be in the same school.

In the schools that were participating in the project, only 9 of the 21 possible TLCs were making reasonable progress (defined as a rating of 3 or 4 in table 2). A more careful analysis of the facilitator logs indicates that the major cause of the poor progress made by the 12 TLCs making slow or no progress (defined as a rating of 2 or 1 in table 2) was that the teachers had not been given time to meet (this was also the case for some of the more successful TLCs, who had decided to commit their own personal time to these meetings because of their interest in, and commitment to, the project). Where time within their contracted hours had been made available for members of TLCs to meet, the TLCs are going well, with considerable enthusiasm for the meetings, and in particular, the teachers appear to value the opportunity to talk about practice in a structured way.

The difficulty that TLCs had in finding time to meet is, at first sight, rather surprising. While none of the schools in Cannington is on the list of approximately 600 schools in England designated by the government as "National Challenge" schools for their failure to achieve the key benchmark of 30% of their students achieving proficiency (grade C or above) in five subjects in the national school-leaving examination for 16 year olds, there is considerable pressure to improve results. Public transport links in Cannington are good, so students can easily travel from one side of the municipality to the other, so parents have a great deal of choice of secondary schools, and this choice is at least informed, if not driven, by examination results at age 16. All the principals say that improving academic outcomes, as measured by success on national examinations taken at ages 16 and 18 is one of their top priorities, and yet despite the research evidence suggesting that formative assessment could make more difference to these academic outcomes than anything else, it appears as if it was difficult for the principals and other school administrators to prioritize the development of formative assessment.

It is even more surprising when one considers that the principals had made a commitment to the program six months before its commencement, had been kept fully informed of the resource requirements necessary for the monthly meetings, and had made a considerable investment in the project by committing an average of 12 teacher days so that teachers could attend the introductory workshop.

The principals of the secondary schools in Cannington remain committed to—and in fact quite positive about—the project, and each of them is looking forward to a "relaunch" of the *Embedding formative assessment* program across the whole school next year, using those who have been involved in the TLCs this year as leaders next year.

Conclusions

The research evidence suggests that when formative assessment practices are integrated into the minute-to-minute and day-by-day classroom activities of teachers, substantial increases in student achievement—of the order of a 70 to 80 percent increase in the speed of learning—are possible, even when outcomes are measured with externally-mandated standardized tests. Indeed, the currently available evidence suggests that there is nothing else that is remotely affordable that is likely to have such a large effect. And while there was clear evidence about how teachers could be supported in developing their use of formative assessment, and that it was feasible in both the US and UK context, how to do so at scale was less clear. In this paper, we have described the development of a pair of products—*Keeping Learning on Track*™ and *Embedding Formative Assessment*—that appear to be effective ways of supporting teachers in their development of formative assessment, even when delivered with little support beyond the resources that can be supplied on a CD-ROM. These resources were designed to be scalable, and to be implemented with minimal additional resources from the school, and within a time allocation that is well within what is routinely spent in schools on administration and bureaucracy. We believe, therefore, that we have made two important steps: clarifying what should be the priority for teacher professional development, and what form that professional development should take.

What we have been surprised to learn, however, is that the third step—actually getting schools to prioritize professional development—appears to be more difficult than either of the first two, and this will be a priority for our future work.

References

Allal, L., & Lopez, L. M. (2005). Formative assessment of learning: a review of publications in French. In J. Looney (Ed.), *Formative assessment: improving learning in secondary classrooms* (pp. 241-264). Paris, France: Organisation for Economic Cooperation and Development.

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

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

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

Black, P., & Harrison, C. (2002). *Science inside the black box: assessment for learning in the science classroom*. London, UK: King's College London Department of Education and Professional Studies.

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

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *Educational evaluation: new roles, new means: the 68th yearbook of the National Society for the Study of Education (part II)* (Vol. 68(2), pp. 26-50). Chicago, IL: University of Chicago Press.

Boekaerts, M. (2006). Self-regulation and effort investment. In K. A. Renninger & I. E.

Sigel (Eds.), *Handbook of child psychology volume 4: child psychology in practice* (6 ed., pp. 345-377). New York, NY: Wiley.

Broadfoot, P. M., Daugherty, R., Gardner, J., Gipps, C. V., Harlen, W., James, M., & Stobart, G. (1999). *Assessment for learning: beyond the black box*. Cambridge, UK: University of Cambridge School of Education.

Brookhart, S. M. (2004). Classroom assessment: tensions and intersections in theory and practice. *Teachers College Record,* **106**(3), 429-458.

Brookhart, S. M. (2007). Expanding views about formative classroom assessment: a review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: theory into practice* (pp. 43-62). New York, NY: Teachers College Press.

Brown, A. L., & Campione, J. C. (1996). Psychological theory and the design of innovative learning environments: on procedures, principles, and systems. In L. Schauble & R. Glaser (Eds.), *Innovations In Learning: New Environments for Education* (pp. 291-292). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ciofalo, J., & Leahy, S. (2006). *Personal action plans: helping to adapt and modify techniques*. Paper presented at the Annual meeting of the American Educational Research Association: San Francisco, CA.

Clymer, J. B., & Wiliam, D. (2006/2007). Improving the way we grade science. *Educational Leadership,* **64**(4), 36-42.

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

Dempster, F. N. (1991). Synthesis of research on reviews and tests. *Educational Leadership,* **48**(7), 71-76.

Dempster, F. N. (1992). Using tests to promote learning: a neglected classroom resource. *Journal of Research and Development in Education,* **25**(4), 213-217.

DuFour, R., DuFour, R., Eaker, R., & Many, T. (2005). *Learning by doing: a handbook for professional learning communities at work*. Bloomington, IL: Solution Tree.

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

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

Gawande, A. (2007). *Better: a surgeon's notes on performance*. London, UK: Profile Books.

Ginsburg, H. P. (2001). *The Mellon Literacy Project: what does it teach us about educational research, practice, and sustainability?* New York, NY: Russell Sage Foundation.

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

Goe, L., & Bridgeman, B. (2006). *Effects of Focus on Standards on academic performance*. Unpublished report. Princeton, NJ: Educational Testing Service.

Grossman, P., Wineburg, S., & Woolworth, S. (2000). *What makes teacher community different from a gathering of teachers?* Seattle, WA: University of Washington Center for the Study of Teaching and Policy.

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

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

Hodgen, J., & Wiliam, D. (2006). *Mathematics inside the black box: assessment for learning in the mathematics classroom*. London, UK: NFER-Nelson.

James, M. (1992). *Assessment for learning* Paper presented at the Annual Conference of the Association for Supervision and Curriculum Development (Assembly session on 'Critique of Reforms in Assessment and Testing in Britain') held at New Orleans, LA.Cambridge, UK: University of Cambridge Institute of Education

Jones, J., & Wiliam, D. (2007). *Modern foreign languages inside the black box: assessment for learning in the modern foreign languages classroom*. London, UK: Granada.

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

Köller, O. (2005). Formative assessment in classrooms: a review of the empirical German literature. In J. Looney (Ed.), *Formative assessment: improving learning in secondary classrooms* (pp. 265-279). Paris, France: Organisation for Economic Cooperation and Development.

Leahy, S., & Wiliam, D. (2009). *Embedding assessment for learning —a professional development pack*. London, UK: Specialist Schools and Academies Trust. https://www.schoolsnetwork.org.uk/pages/default.aspx

Maher, J., & Wiliam, D. (2007). *Keeping learning on track in new teacher induction* Paper presented at the Symposium entitled "Tight but loose: scaling up teacher professional development in diverse contexts" at the annual conference of the American Educational Research Association held at Chicago, IL.

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

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

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

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (Vol. 1, pp. 39-83). Chicago, IL: Rand McNally.

Shepard, L. A. (2007). Formative assessment: caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning* (pp. 279-303). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

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

Thompson, M., & Goe, L. (2008). *Models of effective and scalable teacher professional development* (Research report RR-09-07). Princeton, NJ: Educational Testing Service.

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

Wiliam, D. (2005). *Measuring 'intelligence': what can we learn and how can we move forward?* Paper presented at the Annual meeting of the American Educational Research Association held at Montreal, Canada.

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

Wiliam, D. (2009). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment*. New York, NY: Taylor & Francis.

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

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.

Wylie, E. C., Lyon, C. J., & Goe, L. (2009). *Teacher professional development focused on formative assessment: changing teachers, changing schools* (Vol. RR-09-10). Princeton, NJ: Educational Testing Service.

Wylie, E. C., Lyon, C. J., & Mavronikolas, E. (2008). *Effective and scalable teacher professional development: a report of the formative research and development* (Vol. RR-08-65). Princeton, NJ: Educational Testing Service.

Yeh, S. S. (2006). *Raising student achievement through rapid assessment and test reform*. New York, NY: Teachers College Press.

Address for correspondence: Dylan Wiliam, Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, UK. Telephone: +44 20 7612 6033; Fax: +44 20 7612 6089; Email: d.wiliam@ioe.ac.uk.