Integrating assessment with learning: what will it take to make it work?

Dylan Wiliam and Marnie Thompson, ETS

Introduction

Improving education is a priority for all countries. Increasing the level of educational achievement brings benefits to the individual, such as higher lifetime earnings, and to society as a whole, both in terms of increased economic growth and lower social costs such as health care and criminal justice costs (Gritz & MaCurdy, 1992; Hanushek, 2004; Levin, 1972; Tyler, Murnane, & Willett, 2000). Indeed, the total return on investments in education can be well over $10 for every $1 invested (Schweinhart, et al., 2005). This means that even loosely focused investments in education are likely to be cost-effective. Given public skepticism about such long-term investments, however, and given too the reluctance of local, state, and federal governments to raise taxes, there is a pressing need to find the most cost-effective ways of improving student achievement.

Unfortunately, for many years analysis of school effectiveness focused primarily on easily collected outputs such as average school test scores, and, given the huge differences in the achievement of students entering different schools, between-school differences (what we might call the school effect) appeared to be much larger than within-school differences (such as classroom effects). As disaggregated and longitudinal datasets have become more available, however, it has become possible to look at how much *progress* each student makes (sometimes called “value-added modeling”). Obviously, the estimates of the relative sizes of school and teacher effects will vary

according to the models used, but it is quite common to find that the variability of the teacher effect is much greater than the variability of the school effect (although, as Bryk & Raudenbush, [1988] note, disentangling these two effects is very difficult). In other words, it seems to matter more which teachers you get in a particular school than which school you go to. Perhaps more importantly, the effect of having a good teacher appears to be far greater than that of being in a small school or even of being in a small class.

Hanushek (2004) estimates that students taught by a teacher one standard deviation better than the average—i.e., at the 84th percentile of effectiveness—will achieve 0.22 standard deviations higher than those taught by the average teacher. For example, students taught by a teacher at the 95th percentile of effectiveness will achieve 0.36 standard deviations higher than those taught by average teachers. Because the annual increase in average achievement found in NAEP studies is approximately one-third of a standard deviation (see, for example, NAEP, 2006), this means that students taught by one of the best teachers will learn in six months what students taught by an average teacher would take a year to learn.

This kind of effect is much greater than is found in most class-size reduction studies. For example, Jepsen and Rivkin (2002) found that reducing elementary school class size by ten students would increase the proportion of students passing typical mathematics and reading tests by 4% and 3% respectively, equivalent to a standardized effect size of approximately 0.1 standard deviations for mathematics, and 0.075 for reading. In other words, increasing teacher quality by one standard deviation would produce between two and three times the increase in student achievement yielded by a reducing class-size by ten students. The challenge of improving student achievement at

reasonable cost therefore effectively reduces to a labor force problem with two possible solutions: replacing the teachers we have with better teachers, or improving the teachers we have.

Hanushek (2004) shows that over time even small changes in the way that teachers are hired can lead to large improvements in the quality of the teaching force. Keeping the quality of the teaching force as it is requires hiring at the 50th percentile—this way, the new teachers are just as good as the ones who are leaving. Hiring instead at the 58th percentile would, in 30 years, increase student achievement by 1 standard deviation (and, incidentally, add 10% to per capita GDP). There is however, little evidence that this would be possible. Some have argued that there are large numbers of prospective teachers who would be effective practitioners but who are deterred by burdensome requirements for certification and/or inadequate remuneration (see, e.g., Hess, Rotherham, & Walsh, 2004). Evidence from recent studies suggests, however, that teachers admitted via alternate routes are no more effective than those who follow traditional certification paths (Darling-Hammond, Holtzman, Gatlin, & Vasquez Heilig, 2005), and there is no evidence yet that raising teacher pay attracts better teachers or results in improvements in student learning. In other words, even if one were motivated solely by economic considerations, any significant improvement in educational outcomes will require developing the capability of the existing workforce rather than looking for ways of replacing it—what might be called the "love the one you're with" approach.

Fifteen or twenty years ago, this would have resulted in a gloomy prognosis. There was little if any evidence that the quality of teachers could be improved through teacher professional development, and certainly not at scale. Indeed, there was a

widespread belief that teacher professional development had simply failed to "deliver the goods."

> "Nothing has promised so much and has been so frustratingly wasteful as the thousands of workshops and conferences that led to no significant change in practice when teachers returned to their classrooms" (Fullan, 1991, p. 315).

Within the last few years, however, a clearer picture of the features of effective teacher professional development has begun to emerge. First, teacher professional development needs to attend to both *process* and *content* elements (Reeves, McCall, & MacGilchrist, 2001; Wilson, & Berne, 1999). On the process side, professional development is more effective when it is related to the local circumstances in which the teachers operate (Cobb, McClain, Lamberg, & Dean, 2003), takes place over a period of time rather than being in the form of one-day workshops (Cohen & Hill, 1998), and involves the teacher in active, collective participation (Garet, Birman, Porter, Desimone, & Herman, 1999). It is important to note, however, that some foci for teacher professional development are more productive than others. In particular, professional development is more effective when it has a focus on deepening teachers' knowledge of the content they are to teach, the possible responses of students, and strategies that can be utilized to build on these (Supovitz, 2001). There are many approaches that satisfy both these process and content considerations, such as lesson study (Fernandez & Yoshida, 2004), but in this chapter we will argue that a focus on the use of assessment in support of learning, developed through teacher learning communities, promises not only the largest

potential gains in student achievement, but also provides a model for teacher professional development that can be implemented effectively at scale.

In the following sections, we outline the research on the use of assessment to support learning, sometimes termed formative assessment or assessment for learning, and how the key ideas of formative assessment can be integrated within the broader theoretical framework of the regulation of learning processes. We then summarize briefly the research on teacher learning communities (TLCs) as a powerful mechanism for teacher change, and show how TLCs are perhaps uniquely suited to improving teachers' capabilities in using assessment in the service of learning.

Formative assessment

In 1967, Michael Scriven proposed the use of the terms "formative" and "summative" to distinguish between different roles that evaluation might play. On the one hand, he pointed out that evaluation "may have a role in the on-going improvement of the curriculum" (Scriven, 1967, p. 41) while on the other, evaluation "may serve to enable administrators to decide whether the entire finished curriculum, refined by use of the evaluation process in its first role, represents a sufficiently significant advance on the available alternatives to justify the expense of adoption by a school system" (pp. 41-42). He then proposed "to use the terms 'formative' and 'summative' evaluation to qualify evaluation in these roles" (p.43).

Two years later, Benjamin Bloom (1969, p. 48) applied the same distinction to classroom tests:

> Quite in contrast is the use of "formative evaluation" to provide feedback and correctives at each stage in the teaching-learning process. By formative evaluation we mean evaluation by brief tests used by teachers and students as aids in the learning process. While such tests may be graded and used as part of the judging and classificatory function of evaluation, we see much more effective use of formative evaluation if it is separated from the grading process and used primarily as an aid to teaching.

And there, for a while, things rested. Everyone agreed that formative assessment was “a good thing”; people occasionally used the terms “formative” and “summative” to denote different kinds of assessments, and, when asked, teachers typically said that they did indeed use the results of their assessments to make instructional decisions. The kinds of decisions that teachers made, and the kinds of evidence that informed them were rarely examined however. When the evidence about teachers’ practice did begin to emerge, it was clear that the way that teachers actually used assessment evidence to inform instruction was rarely in the way that Bloom and others had envisaged.

Stiggins and Bridgeford (1985) found that although many teachers created their own assessments, “in at least a third of the structured performance assessment created by these teachers, important assessment procedures appeared not to be followed” (p. 282) and “in an average of 40% of the structured performance assessments, teachers rely on mental record-keeping” (p. 283). A few years later, two substantial review articles, one by Natriello (1987) and the other by Crooks (1988), provided clear evidence that classroom evaluation practices had substantial impact on students and their learning, although the impact was rarely beneficial. Natriello’s review used a model of the assessment cycle, beginning with purposes; and moving on to the setting of tasks, criteria, and standards; evaluating performance and providing feedback and then discussing the

impact of these evaluation processes on students. His most significant point was that the vast majority of the research he cited was largely irrelevant because of weak theorization, which resulted in the conflation of key distinctions (e.g., the quality and quantity of feedback).

Crooks' paper had a narrower focus—the impact of evaluation practices on students. He concluded that the summative function of assessment has been too dominant and that more emphasis should be given to the potential of classroom assessments to assist learning. Most importantly, assessments should emphasize the skills, knowledge, and attitudes regarded as most important, not just those that are easy to assess. The difficulty of reviewing relevant research in this area was highlighted by Black and Wiliam (1998a), in their synthesis of research published since the Natriello and Crooks reviews. The two earlier papers had cited 91 and 241 references respectively, and yet only 9 references were common to both papers. In their own research, Black and Wiliam found that electronic searches based on keywords either generated far too many irrelevant sources, or omitted key papers. In the end, they resorted to manual searches of each issue between 1987 and 1997 of 76 of the journals considered most likely to contain relevant research. Black and Wiliam's review (which cited 250 studies) found that effective use of classroom assessment yielded improvements in student achievement between 0.4 and 0.7 standard deviations. A more recent review focusing on studies in higher education (Nyquist, 2003) found similar results.

Thirty-five years ago, Bloom had suggested that "evaluation in relation to the process of learning and teaching can have strong positive effects on the actual learning of students as well as on their motivation for the learning and their self-concept in relation to school learning. … [E]valuation which is directly related to the teaching-learning process as it unfolds can have highly beneficial effects on the learning of students, the instructional process of teachers, and the use of instructional materials by teachers and learners" (Bloom, 1969, p.50). At the time, Bloom cited no evidence in support of this claim, but it is probably safe to conclude that the question has now been settled: attention to classroom assessment practices can indeed have a substantial impact on student achievement. What is less clear is what exactly constitutes effective classroom assessment, and how the gains in student achievement that the research shows are possible can be achieved at scale. These two issues are the main focus of this chapter.

The nature and purpose of assessments

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

- supporting learning (formative)

- certifying the achievements, or potential of individuals (summative)
- evaluating the quality of educational institutions or programs (evaluative)

Through a series of historical contingencies, we have arrived at a situation in many countries in which the *circumstances* of the assessments have become conflated with the *purposes* of the assessment (Black & Wiliam, 2004a). So, for example, it is often widely assumed that the role of classroom assessment should be limited to supporting learning and that all assessments with which we can hold educational institutions to account must be conducted by an external agency, even though in some countries, this is not the case (Black & Wiliam, 2005a).

In broad terms, moving from formative through summative to evaluative functions of assessment requires data at increasing levels of aggregation, from the individual to the institution; and from specifics of particular skills and weaknesses to generalities about overall levels of performance. (Although, of course, evaluative data may still be disaggregated in order to identify specific sub-groups in the population that are not making progress, or to identify particular weaknesses in students' performance in specific areas, as is the case in France—see Black & Wiliam, 2005a.) It is also clear, however, that the different functions that assessments may serve are not easy to reconcile. For example, when information about student performance on standardized tests is used to hold schools and districts accountable, there is pressure on teachers to raise student performance even if this is at the expense of a narrowed curriculum. As has been shown by the work of Linn and others (see, for example, Linn, 1994), increases in test scores may indicate only that students are better at doing the specific items in the test. In such

cases, the test score provides little information about the aspects of achievement not specifically tested.

For similar reasons, it has been argued that the uses of assessment to support learning and to certify the achievements of individuals are so fundamentally in tension that the same assessments cannot serve both functions adequately (Torrance, 1993). Elsewhere, in a series of papers summarized in Newton (2003) and in Wiliam (2003a), one of us has sketched out how an assessment system might be designed to serve all three functions reasonably well. There are, of course, particular difficulties in implementing such systems where there is a history of students being required to take tests that have little or no consequences for them, as is the case in the United States. For the purposes of this chapter, however, the crucial point is that whatever methods are used for assessing student achievement for summative purposes, assessment still has a role to play in supporting learning. Whether the assessment of student performance for purposes of selection and certification, or for evaluation, is conducted through teacher judgment, external assessments, or some combination of the two, classroom assessment must first be designed to support learning (see Black & Wiliam, 2004b, for a more detailed argument on this point). The remainder of this chapter considers further how this might be done.

What is formative assessment?

In the United States, the term “formative assessment” is often used to describe assessments that are used to provide information on the likely performance of students on state-mandated tests—a usage that might better be described as “early-warning

summative." In other contexts, the term is used to describe any feedback given to students, no matter what use is made of it, such as telling students which items they got correct and incorrect (sometimes called "knowledge of results"). These kinds of usages suggest that the distinction between "formative" and "summative" applies to the assessments themselves, but because the same assessment can be used both formatively and summatively, as both Scriven and Bloom realized, it suggests that these terms are more usefully applied to the use to which the information arising from assessments is put.

In some contexts, assessments that are used to support learning are described under the broad heading "assessment for learning" (in contrast to "assessment *of* learning"). This does suggest a process, rather than being a description of the nature of the assessment itself, but the danger here is that the focus is placed on the intention behind the use of the assessment, rather than action that actually takes place (Wiliam & Black, 1996). Many writers use the terms "assessment for learning" and "formative assessment" interchangeably, but Black, Harrison, Lee, Marshall, and Wiliam (2004, p. 8) distinguish between the two as follows:

> Assessment for learning is any assessment for which the first priority in its design and practice is to serve the purpose of promoting pupils' learning. It thus differs from assessment designed primarily to serve the purposes of accountability, or of ranking, or of certifying competence. An assessment activity can help learning if it provides information to be used as feedback, by teachers, and by their pupils, in assessing themselves and each other, to modify the teaching and learning activities in which they are engaged. Such assessment becomes "formative assessment" when the evidence is actually used to adapt the teaching work to meet learning needs.

For the purpose of this chapter, then, the qualifier "formative" will refer not to an assessment, nor even to the purpose of an assessment, but to the function it actually serves. An assessment is formative to the extent that information from the assessment is fed back within the system and actually used to improve the performance of the system in some way (i.e., the assessment *forms* the direction of the improvement).

So, for example, if a student is told that she needs to work harder, and does work harder as a result, and consequently does indeed make improvements in her performance, this would *not* be formative. The feedback would be *causal*, in that it did trigger the improvement in performance, but not *formative*, because decisions about *how* to "work harder" were left to the student. Telling students to "Give more detail" might be formative, but only if the students knew what giving more detail meant (which is unlikely, because if they knew what detail was required, they would probably have provided it on the first occasion). Similarly, a "formative assessment" that predicts which students are likely to fail the forthcoming state-mandated test is not formative unless the information from the test can be used to improve the quality of the learning within the system. To be formative, feedback needs to contain an implicit or explicit recipe for future action. Sometimes this recipe will be explicit, for example when the feedback identifies specific activities the student is to undertake. At other times, the recipe may be implicit, such as those cases where the teacher has created a culture in the classroom whereby students know they must incorporate any feedback from the teacher into future drafts. Where the teacher relies on implicit recipes, it is of course incumbent on the teacher to check that the students' understanding of the classroom culture is the same as the teacher's.

Another way of thinking about the distinction being made here is in terms of monitoring assessment, diagnostic assessment, and formative assessment. An assessment *monitors* learning to the extent that it provides information about whether the student, class, school or system is learning or not; it is *diagnostic* to the extent that it provides information about what is going wrong; and it is *formative* to the extent that it provides information about what to do about it. A sporting metaphor may be helpful here. Consider a young fast-pitch softballer who has an earned-run average of 10 (for readers who know nothing about softball, that's not good). This is the *monitoring* assessment. Analysis of what she is doing shows that she is trying to pitch a rising fastball (i.e., one that actually rises as it gets near the plate, due to the back-spin applied), but that this pitch is not rising, and therefore ending up as an ordinary fastball in the middle of the strike zone, which is very easy for the batter to hit. This is the *diagnostic* assessment, but it is of little help to the pitcher, because she already knows that her rising fastball is not rising, and that's why she is giving up a lot of runs. If a pitching coach is able to see that she is not dropping her pitching shoulder sufficiently to allow her to deliver the pitch from below the knee, then this assessment has the potential to be not just diagnostic, but *formative*. It provides the athlete with some concrete actions she can undertake in order to improve. This use of formative recalls the original meaning of the term. In the same way that an individual's formative experiences are the experiences that shape the individual, formative assessments are those that shape learning. The important point is that not all diagnoses are *instructionally tractable*—an assessment can accurately diagnose what needs attention without indicating what needs to be done to address the issue.

This was implicit in Ramaprasad's (1983) definition of feedback. According to Ramaprasad, a defining feature of feedback was that it had an impact on the performance on the system. Information that did not have the capacity to improve the performance of the system was not feedback.

> "Feedback is information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way" (Ramaprasad, 1983, p. 4).

In this view, formative assessments (and feedback, in Ramaprasad's terminology) cannot be separated from their instructional consequences, and assessments are formative only to the extent that they impact learning (for an extended discussion on consequences as the key part of the validity of formative assessments, see Wiliam & Black, 1996). The other important feature of Ramaprasad's definition is that it draws attention to three key instructional processes:

- Establishing where the learners are in their learning
- Establishing where they are going
- Establishing what needs to be done to get them there

Traditionally, this may have been seen as the teacher's job, but we need also to take account of the role that the learners themselves, and their peers, play in these processes. The framework shown in Table 1 provides a way of thinking about the key strategies involved in formative assessment. Table 1 could be extended to include schools, districts, or systems. Because the stance taken in this chapter is that ultimately, assessment must

feed into actions in the classroom in order to affect learning, this simplification seems reasonable.

This framework suggests that formative assessment, or assessment for learning (AfL) can be conceptualized as consisting of five key strategies and one "big idea." The five key strategies are

1. Clarifying and sharing learning intentions and criteria for success;
2. Engineering effective classroom discussions, questions, and learning tasks;
3. Providing feedback that moves learners forward;
4. Activating students as instructional resources for one another; and
5. Activating students as the owners of their own learning.

The "big idea" is that evidence about student learning is used to adjust instruction to better meet student needs—in other words that teaching is *adaptive* to the student's learning needs.

Table 1

The five key strategies of formative assessment

	Where the learner is going	Where the learner is right now	How to get there
Teacher	Clarifying learning intentions and criteria for success	Engineering effective classroom discussions, questions, and learning tasks that elicit evidence of learning	Providing feedback that moves learners forward

Peer	Understanding learning intentions and criteria for success	Activating students as instructional resources for one another
Learner	Understanding learning intentions and criteria for success	Activating students as the owners of their own learning

Details of how teachers have used these strategies to implement assessment for learning in their classrooms can be found in Leahy, Lyon, Thompson, and Wiliam (2005). In the remainder of this chapter we discuss how formative assessment may be integrated theoretically with instructional design, and implemented in classrooms.

Formative assessment and the regulation of learning

Although the starting point for work on formative assessment was the relatively simple idea of feedback, the formulation above presents rather a complex picture of formative assessment, and the ways in which the elements within Table 1 relate to each other are not straightforward. All the elements in Table 1 can, however, be integrated within the more general theoretical framework of the *regulation of learning processes* as suggested by Perrenoud (1991, 1998). The word "regulation" has an unfortunate connotation in English, stemming from the idea of "rules and regulations." In French, there are two ways to translate the word regulation—règlement and régulation. The former has the

connotation of rules and regulations, while the latter connotes adjustment, for example in the way that a thermostat regulates the temperature of a room. It is the latter sense that the word is used in the idea of the regulation of learning, and although the term regulation is not ideal to describe this sense in English, in this chapter we will continue its use, not least because of the absence of a suitable alternative.

Within such a framework, the actions of the teacher, the learners, and the context of the classroom are all evaluated with respect to guiding the learning towards the intended goal. In this context, it is important to note that teachers do not create learning; only learners can create learning. In the past, this has resulted in calls for a shift in the role of the teacher from the "sage on the stage" to the "guide on the side." The danger with such a characterization is that it is often interpreted as relieving the teacher of responsibility for ensuring that learning takes place. What we propose here is that the teacher be regarded as responsible for "engineering" a learning environment, both in its design and its operation.

The key features of an effective learning environment are that it creates student engagement and that it is well-regulated. As a growing body of research on cognitive development shows, the level of engagement in cognitively challenging environments influences not only achievement, but also IQ itself (Dickens & Flynn, 2001; Mercer, Dawes, Wegerif, & Sams, 2004). As well as creating engagement, effective learning environments need to be designed so that, as far as possible, they afford or scaffold the learning that is intended (*proactive* regulation). In addition, if the intended learning is not occurring, then this becomes apparent, so that appropriate adjustments may be made (*interactive* regulation).

It is important to distinguish between the regulation of the activity in which the student engages and the regulation of the learning that results. Most teachers appear to be quite skilled at the former, but have only a hazy idea of the learning that results. For example, when asked, “What are your learning intentions for this lesson,” many teachers reply by saying things like, “I’m going to have them describe a friend” conflating the learning intention with the activity (Clarke, 2003). In a way, this is understandable, because only the activities can be manipulated directly. Nevertheless, it is clear that in teachers who have developed their formative assessment practices, there is a strong shift in emphasis away from regulating the activities in which students engage, and towards the learning that results (Black, Harrison, Lee, Marshall, & Wiliam, 2003).

Proactive regulation is achieved “upstream” of the lesson itself (i.e., before the lesson begins), through the setting up of “didactical situations” (Brousseau, 1984). The regulation can be unmediated within such didactical situations, when, for example, a teacher “does not intervene in person, but puts in place a ‘metacognitive culture,’ mutual forms of teaching and the organisation of regulation of learning processes run by technologies or incorporated into classroom organisation and management” (Perrenoud, 1998, p. 100). For example, a teacher’s decision to use realistic contexts in the mathematics classroom can provide a source of regulation, because then students can determine the reasonableness of their answers. If students calculate that the average cost per slice of pizza (say) is $200, provided they are genuinely engaged in the activity, they will know that this solution is unreasonable, and so the use of realistic settings provides a “self-checking” mechanism. Similarly, if a teacher spends time creating a culture in the

classroom in which students are used to consulting and supporting each other in productive ways, then this contributes to keeping the learning "on track."

On the other hand, the didactical situation may be set up so that the regulation is achieved through the mediation of the teacher—*interactive* regulation—when the teacher, in planning the lesson, creates questions, prompts or activities that evoke responses from the students that the teacher can use to determine the progress of the learning, and if necessary, to make adjustments. Sometimes, these questions or prompts will be open-ended questions requiring higher-order thinking—indeed such questions are essential to creating learning environments that create student engagement. But it also important to note that closed questions have a role here too. For example, questions like "Is calculus exact or approximate?", "What is the pH of 10 molar NaOH?", or, "Would your mass be the same on the moon?"are all closed questions, but are valuable because they frequently reveal student conceptions different from those intended by the teacher. Many students believe that calculus is approximate because δx approaches zero, but is not allowed actually to be zero. The pH of 10 molar NaOH is greater than 14, and thus many students think they must have made an error in their calculations because their use of the standard pH indicator has led them to believe that pH cannot be above 14. And although many students know that mass and weight are different, they do not realize that one's mass would be exactly the same on the moon, even though one's weight would be much less. These questions may not cause thinking, but they do provide the teacher with evidence that can be used to adapt instruction to better meet the students' learning needs.

"Upstream" planning of good questions like those above therefore creates, "downstream", the possibility that the learning activities may change course in light of

the students' responses. These "moments of contingency"—points in the instructional sequence when the instruction can proceed in different directions according to the responses of the students—are at the heart of the regulation of learning.

These moments arise continuously in whole-class teaching, where teachers are constantly having to make sense of students' responses, interpreting them in terms of learning needs, and making appropriate responses. But they also arise when the teacher circulates around the classroom, looking at individual students' work, observing the extent to which the students are "on track." In most teaching of mathematics and science, the regulation of learning will be relatively tight, so that the teacher will attempt to "bring into line" all learners who are not heading towards the particular goal sought by the teacher—in these subjects, the *telos* of learning is generally both highly specific and common to all the students in a class. In contrast, in much teaching in language arts and social studies, the regulation will be much looser. Rather than a single goal, there is likely to be a broad *horizon* of appropriate goals (Marshall, 2004), all of which are acceptable, and the teacher will intervene to bring the learners "into line" only when the trajectory of the learner is radically different from that intended by the teacher. Having said this, where a science class is considering the ethical impact of scientific discoveries, or a math class is pursuing an open-ended mathematical investigation, then the regulation is likely to be more like that in the typical language arts classroom. Conversely, where the language arts teacher is covering the conventions of grammar, the regulation is likely to be more like the typical mathematics or science lesson.

Finally, it is worth noting that there are significant differences in how such information is used. In the United States, the teacher will typically intervene with

individual students where they appear not to be "on track" whereas in Japan, the teacher is far more likely to observe all the students carefully, while walking around the class, and then will select some major issues for discussion with the whole class.

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

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

In terms of formative assessment, there are two salient points about such an exchange. First, little is contingent on the responses of the students, except how long it takes to get on to the next part of the teacher's "script," so there is little scope for "downstream" regulation. The teacher is interested only in getting to the word "equilateral" in order that she can move on, and so all incorrect answers are treated as equivalent. The teacher treats all incorrect responses as equivalent in terms of information content; all the teacher learns is that the students didn't "get it."

The second point is that the situation that the teacher set up in the first place—the question she chose to ask—has little potential for providing the teacher with useful information about the students' thinking, except, possibly, whether the students can recall the word "equilateral." This is typical in situations where the questions that the teacher

uses in whole-class interaction have not been prepared in advance (in other words, when there is little or no proactive regulation).

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

A common misconception about assessment—I held by both teachers and administrators—is that testing more frequently makes that testing formative. However, when we think of formative assessment within the framework of regulation of learning, we can see that frequency alone does not guarantee that the information is used to regulate teaching or learning. Frequent assessment can identify students who are not making as much progress as expected (whether this expectation is based on some notion of "ability," prior achievement, or external demands made by the state). But frequent summative testing—we might call this micro-summative—is not formative unless the

information that the tests yield is used in some way to modify instruction (see next section).

System responsiveness and time-frames

Although the examples given above have focused on the classroom, it is important to note that assessments can also be formative at the level of the school, district, and state provided the assessments help to regulate learning. A key issue in the design of assessment systems, if they are to function formatively as well as summatively, is the extent to which the system can respond in a timely manner to the information made available. Feedback loops need to be designed taking account of the responsiveness of the system to the actions that can be used to improve its performance. The less responsive the system, the longer the feedback loops need to be for the system to be able to react appropriately.

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

information generated were used to improve the performance of the system—and if the data from the assessment actually helped to form the direction of the action taken.

For an individual teacher, the feedback loops can be considerably shorter. A teacher might look through the same students' responses to a "trial run" of a state test and re-plan the topics that she is going to teach in the time remaining until the test. Such a test would be useful as little as a week or two before the state-mandated test, as long as there is time to use the information to re-direct the teaching. Again this assessment would be formative as long as the information from the test was actually used to adapt the teaching, and in particular, not only telling the teacher which topics need to be re-taught, but also to suggest what kinds of re-teaching might produce better results.

The building-in of time for responses is a central feature of much elementary and middle school teaching in Japan. A teaching unit is typically allocated 14 lessons, but the content usually occupies only 10 or 11 of the lessons, allowing time for a short test to be given in the 12th lesson, and for the teacher to use lessons 13 and 14 to re-teach aspects of the unit that were not well-understood.

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

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

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

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

All this suggests that the conflicting uses of the term "formative assessment" can be reconciled by recognizing that virtually any assessment can be formative, provided it

is used to make instructional adjustments and that a crucial difference between different assessments is the length of the adjustment cycle. A terminology for the different lengths of cycles is given in table 2.

Table 2

Cycle times for formative assessment

Type	Focus	Length
Long-cycle	Across instructional units, quarters, semesters, years	More than four weeks
Medium-cycle	Between lessons or units	One day to four weeks
Short-cycle	Within a single lesson	Five seconds to one hour

Putting it into practice

No matter how elegantly we formulate our ideas about formative assessment, they will be moot unless we can find ways of supporting teachers in incorporating more attention to assessment in their own practice. There are, of course, other ways that educational research can influence practice, such as through the design of curricula and textbooks, although as Clements (2002) notes, these impacts are generally small. If educational research is to have any lasting impact on practice, it must be taken up and used by practitioners. Traditionally, researchers have engaged in a process of "disseminating" their work to teachers, or engaging in "knowledge transfer." Both of these metaphors have some utility, but they suggest that all researchers need to do is to "share the results" (English, Jones, Lesh, Tirosh, & Bussi, 2002, p. 805) of their research with practitioners and the findings will somehow be used.

The emerging research on expertise shows, however, that the process of "knowledge transfer" cannot be one of providing instructions to novices or other non-expert teachers in the hope that they will get better (see Wiliam, 2003b for more on this point). This is because, put simply, all research findings are generalizations and as such are either too general to be useful, or too specific to be universally applicable. For example, research on feedback, such as the work of Kluger and DeNisi (1996), suggests that task-involving feedback is to be preferred to ego-involving feedback, but what the teacher needs to know is, "Can I say, 'Well done' to this student, now?" Put crudely, such generalizations underdetermine action.

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

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

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

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

There are, of course, many professional development structures that would be consistent with the emerging research base, but professional learning communities, or teacher learning communities (TLCs), as advocated in the *Standards for Staff Development* of the National Staff Development Council (2001), appear to provide the most appropriate vehicle for this work (Borko, 1997, 2004; Borko, Mayfield, Marion, Flexer, & Cumbo, 1997; Elmore, 2002; Kazemi & Franke, 2003; McLaughlin & Talbert, 1993; Putnam & Borko, 2000; Sandoval, Deneroff, & Franke, 2002).

There are several reasons that TLCs are particularly appropriate for the development of teacher expertise in formative assessment. First, formative assessment depends upon a high level of professional judgment on the part of teachers, so it is appropriate to build professional development around a teacher-as-local-expert model. Second, school-embedded TLCs are sustained over time, allowing change to occur developmentally. Third, TLCs are a non-threatening venue allowing teachers to notice

weaknesses in their content knowledge and get help with these deficiencies—in discussing a formative assessment practice that revolves around specific content (e.g., by examining student work that reveals student misconceptions), teachers often confront gaps in their own subject-matter knowledge, which can be remedied in conversations with their colleagues.

Fourth, TLCs are embedded in the day-to-day realities of teachers' classrooms and schools, and thus provide a time and place where teachers can hear real-life stories from colleagues that show the benefits of adopting these techniques in situations similar to their own. These stories provide "existence proofs" that these kinds of changes are feasible with the exact kinds of students that a teacher has in his or her classroom—which contradicts the common lament, "Well, that's all well and good for teachers at *those* schools, but that won't work here with the kinds of students we get at this school." Without this kind of local reassurance, furthermore, there is little chance teachers will risk upsetting the prevailing "classroom contract" (Brousseau, 1997, p. 31). Although limiting, the old contract at least allows teachers to maintain some form of order and matches the expectations of most principals and colleagues. As teachers adjust their practice, they are risking both disorder and less-than-accomplished performance on the part of their students and themselves. Being a member of a community of teacher-learners engaged together in a change process provides the support teachers need to take such risks.

Fifth, and perhaps most importantly, TLCs allow us to address a fundamental limitation of the formative assessment intervention, which is its (perhaps paradoxical) generality and specificity. The five formative assessment strategies that we identified

above are quite general—we have seen each of them in use in pre-kindergarten classes, in graduate-level studies, and at every level in between, and across all subjects. Yet implementing them effectively makes significant demands on subject knowledge. Teachers need good content knowledge to ask good questions, to interpret the responses of their students, and to provide appropriate feedback that focuses on what to do to improve. A less obvious need for subject-matter knowledge is that teachers need a good overview of the subject matter in order to be clear about what the "big ideas" are in a particular domain so that these are given greater emphasis. Thus TLCs provide a forum for supporting teachers in converting the broad formative assessment strategies into "lived" practices within their classrooms.

Creating the conditions and support mechanisms for establishing and sustaining TLCs in schools is non-trivial, and, indeed, challenges long-held structures and assumptions about the nature of teachers' work, teacher learning, and how time should be spent in school. It is our contention, though, that getting teachers to integrate assessment for learning directly into their practice will have such positive effects on student learning, that it is worth investing research and development efforts in support of developing scalable TLC models.

Conclusion

In this talk, we have argued that the terms formative and summative apply not to assessments themselves, but to the functions they serve, and as a result, it is possible for the same assessment to be both formative and summative. Assessment is formative when the information arising from the assessment is fed back within the system and is actually used to improve the performance of the system. Assessment is formative for individuals

when they use the feedback from the assessment to improve their learning. Assessment is formative for teachers when the outcomes from the assessment, appropriately interpreted, help them improve their teaching, either on specific topics, or generally. Assessments are formative for schools and districts if the information generated can be interpreted and acted upon in such a way as to improve the quality of learning within the schools and districts.

The view of assessment presented here involves a shift from quality control in learning to quality assurance in learning. Rather than teaching students, and then, at the end of the teaching, finding out what has been learned, it seems obvious that what we should do is to assess the progress of learning whilst it is happening, so that we can adjust the teaching if things are not working. In order to achieve this, the length of the cycle from evidence to action must be designed taking into account the responsiveness of the system. Some feedback loops, such as those in the classroom, will be only seconds long, while others, such as those involving districts or state systems will last months, or even years.

More generally, we have suggested that formative assessment be considered as a key component of well-regulated learning environments. From this perspective, the task of the teacher is to not necessarily to teach, but rather to engineer and regulate situations in which students learn effectively. One way to do this is to design the environment so that the regulation is embedded within features of the environment. Alternatively, when the regulation is undertaken through the teacher's mediation, it is necessary to build opportunities for such mediation into the instructional sequence by designing in episodes

that will elicit students' thinking (proactive regulation) and to use the evidence from these probes to modify the instruction (interactive regulation).

Work with teachers to date suggests that the development of teachers' formative assessment practices through the use of teacher learning communities is manageable and relatively inexpensive to implement. The changes are slow to take effect, however, and it is not yet clear how faithfully the model used here can be scaled up. We have begun to explore what, exactly, changes when teachers develop formative assessment (Black & Wiliam, 2005b), but much more remains to be done. In particular, we do not know whether some of the five key strategies identified above have greater leverage than others, both for promoting professional development and for increasing student achievement. Nevertheless, we believe that there are reasons to be optimistic. Perhaps one day we will not talk about "integrating assessment with learning" because the distinction between the two will be meaningless.

Acknowledgements

We are grateful to Carol Dwyer, Laura Goe, and Siobhan Leahy for comments on an earlier version of this chapter.

References

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

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

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

Black, P. J., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London, UK: King's College London School of Education.

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

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

Black, P., & Wiliam, D. (2005a). Lessons from around the world: How policies, politics and cultures constrain and afford assessment practices. *Curriculum Journal, 16*(2), 249-261.

Black, P., & Wiliam, D. (2005b). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning*. London: Sage.

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *Educational evaluation: New roles, new means. The 63rd yearbook of the National*

Society for the Study of Education, part 2. (Vol. 69, pp. 26-50). Chicago: University of Chicago Press.

Borko, H. (1997). New forms of classroom assessment: Implications for staff development. *Theory into practice, 36*(4), 231-238.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8), 3-15.

Borko, H., Mayfield, V., Marion, S. F., Flexer, R. J., & Cumbo, K. (1997). *Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development* (Vol. 423). Boulder: University of Colorado, Boulder Center for Research on Evaluations, Standards and Student Testing (CRESST).

Brousseau, G. (1984). The crucial role of the didactical contract in the analysis and construction of situations in teaching and learning mathematics. In H. G. Steiner (Ed.), *Theory of mathematics education*, pp. 110-119. Occasionnel Paper 54, Bielefeld, Germany, University of Bielefeld, Institut für Didaktik der Mathematik.

Brousseau, G. (1997). *Theory of didactical situations in mathematics* (N. Balacheff, M. Cooper, R. Sutherland, & V. Warfield, Trans.). Dordrecht, Netherlands: Kluwer.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*, 65-108.

Clarke, S. (2003). *Enriching feedback in the primary classroom*. London: Hodder & Stoughton.

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

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

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

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

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Vasquez Heilig, J. (2005). *Does teacher preparation matter? Evidence about teacher certification, teach for America, and teacher effectiveness.* Stanford, CA: Stanford University School of Education.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review, 108,* 346-369.

Elmore, R. F. (2002). *Bridging the gap between standards and achievement: Report on the imperative for professional development in education*. Washington, DC: Albert Shanker Institute.

English, L. D., Jones, G., Lesh, R., Tirosh, D., & Bussi, M. B. (2002). Future issues and directions in international mathematics education research. In L. D. English (Ed.),

Handbook of international research in mathematics education (pp. 787-812). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Fullan, M. (1991). *The new meaning of educational change*. London: Cassell.

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

Marshall, B. (2004). Goals or horizons—the conundrum of progression in English: Or a possible way of understanding formative assessment in English. *Curriculum Journal, 15*(2), 101-113.

Gritz, R. M., & MaCurdy, T. (1992). *Participation in low-wage labor markets by young men* (Discussion Paper). Washington, DC: Bureau of Labor Statistics.

Hanushek, E. A. (2004). *Some simple analytics of school quality* (NBER working paper No. W10229). Washington, DC: National Bureau of Economic Research.

Hess, F. M., Rotherham, A. J., & Walsh, K. (Eds.). (2004). *A qualified teacher in every classroom? Appraising old answers and new ideas*. Cambridge, MA: Harvard Education Press.

Jepsen, C., & Rivkin, S. G. (2002). *What is the tradeoff between smaller classes and teacher quality?* (NBER working paper No. 9205). Cambridge, MA: National Bureau of Economic Research.

Kazemi, E., & Franke, M. L. (2003). *Using student work to support professional development in elementary mathematics: A CTP working paper*. Seattle, WA: Center for the Study of Teaching and Policy.

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.

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

Levin, H. M. (1972). The costs of inadequate education. In Select Committee on Equal Educational Opportunity - United States Senate (Ed.), *Toward equal educational opportunity* (pp. 171-186). Washington, DC: U.S. Government Printing Office.

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

Linn, R. L. (1995). *Assessment-based reform: Challenges to educational measurement* (Report No. PIC-ANG1). Princeton, NJ: William H. Angoff Memorial Lecture Series, Educational Testing Service.

McLaughlin, M., & Talbert, J. (1993). *Contexts that matter for teaching and learning: Strategic opportunities for meeting the nation's educational goals.* Palo Alto, CA: Stanford University Center for Research on the Context of Secondary School Teaching.

Mercer, N., Dawes, L., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. *British Educational Research Journal, 30*(3), 359-377.

National Assessment of Educational Progress. (2006). *The nation's report card: Mathematics 2005* (NCES 2006-453). Washington, DC: Institute of Education Sciences.

National Staff Development Council. (2001). *NSDC standards for staff development*. Oxford, OH: National Staff Development Council.

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

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

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

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

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

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*(1), 4-15.

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

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

Sandoval, W., Deneroff, V., & Franke, M. L. (2002). *Teaching, as learning, as inquiry: Moving beyond activity in the analysis of teaching practice.* Paper presented at the 2002 Annual Meeting of the American Educational Research Association, New Orleans, LA.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects: The High/Scope Perry Preschool study through age 40*. Ypsilanti, MI: High/Scope Educational Research Foundation.

Scriven, M. (1967). *The methodology of evaluation* (Vol. 1). Washington, DC: American Educational Research Association.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*(4), 271 - 286.

Supovitz, J. A. (2001). Translating teaching practice into improved student achievement. In S. H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the States* (Vol. 2, pp. 81-98). Chicago: University of Chicago Press.

Torrance, H. (1993). Formative assessment: Some theoretical problems and empirical questions. *Cambridge Journal of Education, 23*(3), 333-343.

Tyler, J. H., Murnane, R. J., & Willett, J. B. (2000). *Estimating the labor market signaling value of the GED* (Research Brief). Boston: National Center for the Study of Adult Learning & Literacy.

Wiliam, D. & Black, P. J. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22*(5), 537-548.

Wiliam, D. (1999, May) *A template for computer-aided diagnostic analyses of test outcome data.* Paper presented at 25th annual conference of the International Association for Educational Assessment, Bled, Slovenia.

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

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

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

Wilson, S. M. & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. In

A. Iran-Nejad, & P. D. Pearson (Eds.), *Review of research in education* (pp. 173-209). Washington, DC: American Educational Research Association.