Formative assessment: getting the focus right

Dylan Wiliam, ETS

Writing in 1967, Michael Scriven suggested two roles that evaluation might play. On the one hand, " it may have a role in the on-going improvement of the curriculum" (Scriven, 1967 p.41) while in another role, " the evaluation process 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 suggested that the same distinction might be applied to the evaluation of student learning—what we today would tend to call "assessment". He acknowledged the traditional role that tests played in judging and classifying students, but noted that there was another role for evaluation:

> 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 (Bloom 1969, p.48)

Explicit in these early uses is that the term "formative" cannot be a property of an assessment. As Bloom makes clear, the same tests could be used for formative or summative uses, although he suggests that the formative use will be less effective if the tests are part of the grading process. The crucial feature of formative evaluations, for both Scriven and Bloom, is that the information is used in some way to make changes. Whether it is a curriculum or student achievement that is being evaluated, the evaluation is formative if the information generated is used to make changes to what would have happened in the absence of such information. In the same way that one's formative experiences are those experiences that shape us as individuals, formative evaluations are those that shape whatever is being evaluated. An assessment of a curriculum is formative if it shapes the development of that curriculum. An assessment of a student is formative if it shapes that student's learning. Assessments are formative, therefore, if and only if something is contingent on their outcome, and the information is actually used to alter what would have happened in the absence of the information.

Of course, the time-scale for this depends on the decisions that need to be made. For example, a science supervisor in a school district may need to plan in the spring the workshops that she will offer to teachers during the summer. She may look at the scores obtained by students in last year's state tests, and find that the levels of performance on some domains is lower, relative to the average for the state, than others. She could then

use this information to plan workshops on the areas where performance was weakest, thus meeting the learning needs of the science teachers in the district in a way that would not have been possible, or at least would have been unlikely, without that information. In this example, the information about the performance of the students in the district was used by the supervisor to adapt her plans for the summer workshops in order to better meet the teachers' learning needs. At the other extreme, a language arts teacher might ask a class of students the following question:

Which of these is a good thesis statement?
A) The typical TV show has 9 violent incidents
B) There is a lot of violence on TV
C) The amount of violence on TV should be reduced
D) Some programs are more violent than others
E) Violence is included in programs to boost ratings
F) Violence on TV is interesting
G) I don't like the violence on TV
H) The essay I am going to write is about violence on TV

and require each student to respond by holding up one (or more) of a set of cards labeled A, B, C, D, E, F, G and H. At this point, the teacher has created a "moment of contingency"—a point in the instructional sequence where the instruction can change direction in light of evidence about the students' achievement, thus allowing her to adapt the instruction to better meet their learning needs. If all students choose option C, the teacher can move on, reasonably confident that all students in her class understand what constitutes a good thesis statement. If most of the students have not answered correctly, then the teacher has created a "teachable moment." If most of the students have chosen incorrectly, she might choose to review the work on thesis statements with the class. But if some have answered correctly while others have not, then she might initiate a class discussion. Moreover, because she knows which students chose which option, she can use this information to guide the discussion more effectively.

These two examples illustrate the extreme ends of a continuum. The science supervisor is engaged in what we might term "long-cycle" formative assessment. The cycles here can be years long. The results of tests that students took in March 2005 might be used to plan workshops for teachers in the summer of 2006, but these could not affect student scores on state tests until the tests taken in March 2007—and these results might not be known until summer 2007. At the other extreme, the language arts teacher is using a cycle that is minutes, if not seconds, in duration—what we might call "short-cycle" formative assessment. The focus of many of the studies in this issue has been somewhere between these two extremes—what we might call "medium-cycle" formative assessment. In table 1 below, the foci and typical durations of these differences, based on Wiliam and Thompson (2006) are given.

What makes an assessment formative, therefore, is not the length of the feedback loop, nor where it takes place, nor who carries it out, nor even who responds. The crucial

feature is that evidence is evoked, interpreted in terms of learning needs, and used to make adjustments to better meet those learning needs.

Type	Focus	Length
Long-cycle	between instructional units	four weeks to one year or more
Medium-cycle	between lessons	one day to two weeks
Short-cycle	within a single lesson	five seconds to one hour

Table 1: Types of formative assessment

All the papers in this special issue highlight these aspects to a greater or lesser extent. The analysis by Ruiz-Primo and Aracelli shows that the teachers who most consistently elicit the right kinds of information (conceptual eliciting questions), who have ways of interpreting the students' responses in terms of learning needs, and who can use this information to adapt their instruction, generate higher levels of student achievement.

The paper by Niemi et al examines the use of representations in mathematics classrooms in terms of a five-step model of the teaching and assessment process. Their model is prescriptive rather than descriptive; while there can be little doubt of the utility of proper content analysis before beginning instruction, there is little evidence that this is what teachers actually do, nor is there much evidence that teachers take much account of students' prior knowledge. Their analysis focuses on the key role of the process of representation, although here representations are restricted to external (rather than, say, cognitive) representations, akin to what Bruner (1996) calls *oeuvres* ("works"). Here again the importance of eliciting the right information (what representations should we get students to think about?), and making sense of their responses, comes through clearly.

In this regard, the "archaeology" of the assessment may be relevant. It is notable that in many of the papers in this volume, and in most of the research reported in recent years, teachers have tried to adapt assessments originally designed for summative purposes (e.g. grading) for formative purposes. In the paper by Gearhart et al., while one teacher did use white-boards to elicit responses from groups that the teacher could use to adapt instruction in real time, she, and the other two teachers, focused more on using more or less formal assessment episodes with rubrics that could support summative inferences. As a result, the evidence collected by the teachers was more useful for describing where students were than for informing how to move them forward.

Similar issues are raised in the paper by Aschbacher and Alonzo. Science notebooks have potential as formative assessments—i.e. they can generate evidence of student achievement that would allow a teacher to adapt her instruction—but this happens only when the prompt used for eliciting evidence is structured appropriately. For students in Mrs Cruz's class, because the task was too highly structured, all we learn is whether the students can follow directions. In contrast, in those classes where too little guidance was given, we get useful information on the understanding of some students, but for other

students, we learn very little. In particular, we do not know whether the lack of evidence is due to lack of understanding of the science, or lack of understanding about what they were being asked to do. In contrast, the structure for the notebook entries in Mrs Perez's class increases the *disclosure* of the task (Wiliam, 1992)—the likelihood that "if they know it, they show it".

The nature of the domain being studied also has profound implications for the kinds of assessments that are likely to be useful. In science, we can often itemize the knowledge we want students to acquire. This makes it relatively straightforward to move from monitoring (is learning taking place?) to diagnosis (what is not being learned?) to action (what to do about it?). In a domain such as reading, however, the cause of the problem is much less clear. Scarborough (2001) shows that skilled reading involves the simultaneous articulation of a large number of skills, including phonological awareness, decoding, sight recognition of familiar words, background knowledge of the subject of the text, vocabulary, knowledge about language structures, inferential skills, and even knowledge of literacy concepts such as genre. The paper by Bailey and Drummond shows that early-years teachers can generally identify which students are struggling, but are less skilled at identifying the causes of the failure to progress. As with the other papers in this issue, the first challenge is not how to interpret the evidence in terms of student learning needs, but in generating the right evidence in the first place.

Taken as a whole, the papers in this issue make important contributions to our understanding of how difficult it is likely to be to improve teachers' use of formative assessment strategies. In particular, they suggest that while the provision of high-quality tools may be a necessary condition, it is certainly not a sufficient condition for the improvement of formative assessment practice. Tools for formative assessment will only improve formative assessment practices if teachers can integrate them into their regular classroom activities. In other words, the task of improving formative assessment is substantially, if not mainly, about teacher professional development.

Fifteen 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).

In recent years, however, we have learned that to be effective professional development needs to attend to both *process* and *content* elements (Reeves, McCall, and MacGilchrist, 2001; Wilson and 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, and Dean, 2003), takes place over a period of time rather than being in the form of one-day workshops (Cohen and Hill, 1998), and involves the teacher in active, collective participation (Garet, Birman, Porter, Desimone, and Herman, 1999). On the content side professional development is more effective when it has a focus on

deepening teachers' knowledge of the content they are to teach, the possible responses of students, and strategies that can be utilized to build on these (Supovitz, 2001).
The creation of teacher learning communities (TLCs) focused on formative assessment appear to show the greatest potential for improving teaching practice and student achievement (Wiliam and Thompson, 2006), but a note of caution is in order here.

As noted above, most of the studies reported in this issue have focused on medium cycle formative assessment, but my own reading of the research (e.g. Black and Wiliam, 1998) suggests that medium-cycle formative assessments have shown only modest impact on student learning. Why this is so is not clear. In some cases, it may be because many of the assessments being used were summative assessments pressed into service for formative purposes, rather than being designed from the outset to be formative (the too easy equation of "formative assessment" with "classroom assessment" may be partly responsible here). It may be that it is just too hard for teachers to use information at the end of a sequence of learning to adapt instruction, due to the pressure from curriculum pacing guides or sequencing charts. Certainly the studies that have shown impact on student learning (e.g. Wiliam, Lee, Harrison & Black, 2004) have tended to be those where the introduction impacted teachers' day-to-day and minute-to-minute classroom practices, either by an explicit focus on short cycle assessment (Leahy, Lyon, Thompson and Wiliam, 2005) or where a focus on medium or long cycle formative assessment was implemented in such a way as to require teachers to change their regular classroom practice. Further work is needed to elaborate the "logic model" of formative assessment to clarify exactly what we believe the interventions are changing, and how much impact they have on student learning.

My final concern in all this is that many, if not most, research efforts on supporting teachers in the use of formative assessment represent a "counsel of perfection." There is a focus on meeting the needs of all students that is laudable, but simply unlikely to be possible in most American classrooms. American teachers are some of the most hard-working in the world, with around 1130 contact-hours per year compared to the OECD average of 803 hours for primary and 674 hours for upper secondary (OECD, 2003). If we are to effect substantial change at scale, we need to focus on the changes that we can produce most easily. Of course, we must not tolerate interventions that exacerbate existing inequality, but the pressing need now is to move teachers to action. As Robert Slavin (1987) remarked in another context, "Do we really know nothing until we know everything?"

References

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

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 II)* (Vol. 69(2), pp. 26-50). Chicago, IL: University of Chicago Press.

Bruner, J. S. (1996). *The culture of education*. Cambridge, MA: Harvard University Press.

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

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

Desimone, L., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, **24**(2), 81-112.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, **38**(4), 914-945.

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

Organisation for Economic Cooperation and Development. (2003). *Education at a glance*. Paris, France: Organisation for Economic Cooperation and Development.

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.

Scarborough, H. (2001). Connecting early language and literacy to later reading (dis)abilities: evidence, theory and practice. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research*. New York, NY: Guilford Press.

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.

Slavin, R. E. (1987). Ability grouping in elementary schools: do we really know nothing until we know everything? *Review of Educational Research*, **57**(3), 347-350.

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. Part 2, pp. 81-98). Chcago, IL: University of Chicago Press.

Wiliam, D. (1992). Some technical issues in assessment: a user's guide. *British Journal for Curriculum and Assessment,* **2**(3), 11-20.

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.

Wiliam, D., & Thompson, M. (2006). Integrating assessment with instruction: what will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

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* (Vol. 24, pp. 173-209). Washington, DC: American Educational Research Association.