Keeping learning on track: classroom assessment and the regulation of learning[i]

To appear in F. K. Lester Jr (Ed.), *Second handbook of mathematics teaching and learning*. Greenwich, CT: Information Age Publishing (2007).

Dylan Wiliam, University of London Institute of Education

Introduction

When teachers are asked how they assess their students, they are likely to cite tests, quizzes, portfolios, projects and other more or less formal methods. When, instead, teachers are asked how they know whether their students have learned something, the responses are typically very different (Dorr-Bremme & Herman, 1986). They mention classroom questions, group activities, discussions, posters, concept maps, and even the expressions on the faces of their students and in fact, the origin of the word assessment (Latin assidere; literally "to sit beside") is much closer to this more informal, meaning. However, the emphasis on assessment as a formal process is pervasive, and mathematics education is no exception.

To a first approximation, then, the research literature on assessment, both generally, and in mathematics education, is almost entirely about the formal methods of assessment, and in particular, tests and examinations. To make matters worse, even when less formal methods of assessments, such as teacher-made tests are discussed, the purpose of an assessment is far more likely to be that of making a determination of a student's existing state of knowledge. Glaser and Silver (1994) observe that, "Aside from teacher-made classroom tests, the integration of assessment and learning as an interacting system has been too little explored" (p. 403). Thus, even when classroom assessment is studied, the emphasis has tended to be on the concordance of such measures with external measures—in other words, on classroom assessment as an alternative to external assessments. As Kilpatrick, Swafford and Findell (2001) note in their survey *Adding it up: helping children learn mathematics*, "Even less attention appears to have been paid to how teachers' assessments might help improve mathematics learning" (p. 40).

The intent of this chapter is to redress the balance, by focusing on the role that assessment can play in supporting learning, rather than just measuring it—sometimes described as a distinction between assessment *for* learning and assessment *of* learning[ii], but two further qualifications are needed here. The first is that assessment can support learning in a variety of ways such as, for example, when students actually learn something while completing an assessment, as emphasized by Sternberg & Williams (1998 p. 10)—a process that might be termed "assessment as learning". As Shavelson, Baxter and Pine (1992) note, "a good assessment makes a good teaching activity, and a good teaching activity makes a good assessment" (p. 22). Pellegrino, Chudowsky, & Glaser (2001) and Shepard, Hammerness, Darling-Hammond, Rust, Snowden, Gordon, Gutierrez & Pacheco (2005) summarize the research on the design of such formal assessments, and Lester, Lambdin and Preston (1997) deal specifically with the possibilities for mathematics.

Accordingly the specific focus of this chapter is not on how teachers and students can use assessment activities to promote learning. It is about assessment as an essentially interactive process, in which the teacher can find out whether what has been taught has been learned, and if not, to do something about it. It is therefore about assessment functioning as a bridge between teaching and learning, helping teachers collect evidence about student achievement in order to adjust instruction to better meet student learning needs, in real time.

The second qualification is that the focus of this chapter is firmly on the learning of mathematics in the mathematics classroom. Effective implementation of the kinds of exemplary practices identified in this chapter will entail careful consideration of a whole range of issues related to classroom management (see for example, Brookhart, 2004), teacher professional development (Wiliam & Thompson, to appear) and educational policy (Looney, 2005). While these issues are clearly important, they are beyond the scope of this chapter.

The importance of the role of assessment in instruction was explicitly recognized in the NCTM's six *Assessment standards for school mathematics* (NCTM, 1995), the second of which states that "Assessment should enhance mathematics learning" (p. 13). The NCTM's Assessment standards also make clear that assessment has a role to play not just in making determinations about whether particular teaching activities were successful, but also in teachers' moment-by-moment decisions making (p. 46)—in other words, that teachers should use assessment to "keep learning on track."

This distinction about the *purpose* of assessment is quite different from the distinction between classroom assessment and external assessments, which is more concerned with where the assessments take place, who sets them and who scores them (Black and Wiliam, 2004a). This is an important qualification, particularly in the USA, where the term classroom assessment is used primarily to mean classroom *summative* assessment.

The next section lays out the reasons for two key assumptions that have been made in the writing of this chapter, namely that assessment is independent both of curriculum and of any particular stance in psychology—in other words, that we can talk about the principles of good assessment without subscribing to a particular view of what should be in the mathematics curriculum, or even to a particular view about what happens when learning takes place. The reader who is prepared to accept these two assertions may comfortably skip the next section without loss of continuity. In subsequent sections, the research on the impact of assessment on learning is reviewed, and it is suggested that the effective use of assessment for learning consists of five key strategies:

1. Clarifying and sharing learning intentions and criteria for success;
2. Engineering effective classroom discussions, questions, and learning tasks that elicit evidence of learning;
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.

In the final section of the chapter, these five strategies are subsumed within a broader theoretical framework, namely the regulation of learning processes, which allows assessment to be integrated with principles of instructional design.

Key assumptions

There are two major challenges in synthesizing the research on assessment *for* learning (or formative assessment as it is sometimes called). The first is to what extent does an adequate account of assessment for learning require agreement about what mathematics students should learn? The second is to what extent does an adequate account of assessment for learning require agreement about what happens when learning takes place? Each of these is discussed in turn below.

What students should learn in mathematics is a highly contested domain. For example, in discussing the widespread preconception that mathematics is about learning to compute, Fuson, Kalchman & Bransford (2005) illustrate their discussion with the following question:

What, approximately, is the sum of 8/9 plus 12/13?

They point out that some people will, sensibly, conclude that the answer is a little less than two, just by observing that the two numbers to be added are each a little less than one. Others, however, will attempt to find the smallest common multiple of the denominators of the two fractions. They comment:

> The point of this example is not that computation should not be taught or is unimportant; indeed, it is often critical to efficient problem solving. But if one believes that mathematics is about problem solving and that computation is a tool for use to that end when it is helpful, then the above problem is viewed not as a "request for computation," but as a problem to be solved that may or may not require computation—and in this case, it does not (Fuson *et al*., p. 220)

For many people involved in mathematics education, estimation skills are at least as important as, and perhaps more important than, computation skills. Almost 30 years ago, Michael Girling defined "numeracy" as "the ability to use a four-function calculator sensibly" (Girling, 1977 p. 4), suggesting that the skill of being able to assess the reasonableness of an answer was much more important than being able to compute it accurately. Others feel that estimation is much less important than a solid grounding in basic mathematical knowledge and skills (Klein, 2003).

These differences become even clearer in more advanced mathematics. For example, many people need to use the standard formula for solving quadratic equations:

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Two things seem important here. The first is that, obviously, this formula needs to be memorized exactly if it is to be of any use. The second is that while some of the people who know this formula would be able to re-create it from scratch by the process of "completing the square," most would not. It seems likely that professionals involved in mathematics education would agree that the person who could derive this formula from scratch has a deeper understanding than someone who can merely reproduce the formula from memory, but that is not to say that simply knowing this formula is not useful. Indeed,

anyone who knows and can apply this formula accurately has what Richard Skemp called "instrumental understanding." This is a different kind of understanding from knowing where the formula comes from and how to derive it (what Skemp called "relational understanding") but it is still a form of understanding (Skemp, 1977). The crucial point here is that different users of mathematics have different needs, and that a well-grounded account of the role of assessment in mathematics education should serve them all.

Building on the work of Raymond Williams (1961), Paul Ernest (1991) identifies five broad purposes for mathematics education:

> Acquiring basic mathematical skills and numeracy and social training in obedience;
> Learning basic skills and learning to solve practical problems with mathematics and information technology;
> Understanding and capability in advanced mathematics, with some appreciation of mathematics;
> Gaining confidence, creativity and self expression through mathematics;
> Empowerment of learners as critical and mathematically literate citizens in society.

The obvious corollary of the fact that there are differences in people's perceptions of the purpose of mathematics is that those with different aims will emphasize different aspects of mathematics. For those who see the first of these five broad aims as the major purpose of mathematics education, they will value mathematics curricula that emphasize this purpose. For those who regard the last of the five as most important, they will value very different curricula. The important point here is that while competing groups can construct arguments to justify their claims (Niss, 1993), these are essentially *value* arguments. In particular, there is no way for adherents of one particular view of the purpose of education to show dissenters that they are wrong. This is why an adequate account of classroom assessment must support any and all of these conflicting views of mathematics education, rather than imposing a certain set of views.

Similar arguments apply to the psychology of mathematics education. For the first half of the twentieth century, the dominant view about what happens when learning takes place was that the individual creates associations between stimuli and responses. These "associationist" views of learning included the behaviorism of Skinner (for example), as well as a range of other views. Associationist models of learning explained some aspects of mathematics learning reasonably well, but were unable to explain other aspects. For example, an associationist analysis of students' errors in learning multiplication facts would indicate that students errors should be random—the result of insufficient reinforcement of particular links in chains of stimulus and response—which accords reasonably well with what we observe in practice. However, in other areas of mathematics learning, students' errors are clearly not random; in fact they are highly predictable (see examples in the section on "eliciting evidence" below). The mounting evidence about the non-random nature of students' errors in mathematics led to the development of "constructivist" approaches to the study of learning, where it is acknowledged that students are active rather than passive in the development of their conceptions (see, for example, von Glasersfeld, 1991). Such theories were much better able to account for the systematicity in student errors, but there were some aspects of learning that constructivist views were unable to explain. For example, it was observed that some adults were able to perform calculations in some contexts (e.g. in supermarkets) but not others, such as classrooms (Lave, Murtaugh & de la Roche, 1984). This idea that learning is often tied to the context in which learning

takes place has proved to be particularly powerful in mathematics education. For example, Boaler (2002) found that in some mathematics classrooms, students were able to use the mathematics they had learned outside school, while in others, students were not. This suggests that theories in psychology are not like theories in, say, physics, where new theories include previous theories as special cases (for example in the way that Einstein's relativistic mechanics includes Newton's non-relativistic mechanics as a special case).

Rather, in psychology, the tendency is for each new theory to be very good at explaining what previous theories did not, but generally not so good at explaining what the previous theories explained well. Many kinds of rote learning are well explained by associationist theories, while the regularities that we see in students errors are better explained by constructivist or information-processing theories, and situated theories explain transfer, or its absence, well. In this sense, each new theory does not replace the preceding theories, but rather complements them. For different views of what mathematics should learn, there may be different views of what happens when learning takes place, although it is important to note that no one view of learning will suffice for even the most specific learning (Sfard, 1998). For example, it might appear at first sight that learning multiplication facts would be primarily a matter of strengthening associations between stimuli and responses, but it has been shown that many students "repair" gaps in their knowledge by using their knowledge of the processes of arithmetic to assemble and combine facts and routines that they can recall (VanLehn, 1990).

For the purpose of this chapter, the important point is that if it is to be useful, an adequate account of classroom assessment cannot dictate what mathematics students should learn nor should it be tied to a single view of what happens when learning takes place.

This chapter will, therefore, as far as possible, avoid putting the assessment cart before the either the curriculum or psychology horse. The stance being taken is that assessment is a powerful servant but a bad master. As soon as assessment considerations are allowed to influence either what is to be learnt, or what it means to learn, we are likely to slip from making the important assessable to making the assessable important.

The purposes of assessment

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 (Black and Wiliam, 2004a). In particular, each of these can be the responsibility of those who teach the students, or, 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 programs or institutions (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 and 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 and 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 and Wiliam, 2005a). However, it is also clear that the different functions that assessments may serve are in tension. The use of data from assessments to hold schools accountable has, in many cases, because of "teaching to the test," rendered the data almost useless for attesting to the qualities of individual students (apart, of course, from those qualities that are tested) or for supporting learning.

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). On the other hand, others have argued that ways must be found to integrate the two (e.g. Shavelson, Black, Wiliam & Coffey, 2003). For the purposes of this chapter, the crucial point is that the use of assessment should support instruction in *any* assessment regime. Whether the assessment is for purposes of selection and certification, or for evaluation, whether it is conducted through teacher judgment, external assessments, or some combination of the two, classroom assessment must first be designed to support learning (see Black and Wiliam, 2004b, for a more detailed argument on this point). The remainder of this chapter considers further how this might be done.

Formative assessment: origins and examples

In 1967, Michael Scriven proposed the use of the terms "formative" and "summative" to distinguish between different roles that evaluation[iii] 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 in another role, 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." (Scriven, 1967, 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.

However, despite Bloom's extension of the term "formative" to apply to the evaluation of individual students (what in this chapter is termed "assessment") as well as the evaluation of programs or institutions, for the next thirty years, the term "formative" was used almost exclusively in the context of program evaluation. Indeed, the index of the previous NCTM handbook of research on mathematics education lists only one mention of the term "formative" and that is in the section on "Evaluation" in the chapter on Research Methods (Romberg, 1992 p. 58).

Nevertheless, while the term "formative" was rarely used to describe teachers' assessment practices, a number of studies investigated the integration of assessment with instruction, the best known of which is probably Cognitively-Guided Instruction (CGI).

In the original CGI project, a group of 21 elementary school teachers participated in a series of workshops over a four year period (an introductory $2\frac{1}{2}$ hour workshop and a 2-day workshop before the beginning of the first school year, fourteen 3-hour workshops during the first year, four $2\frac{1}{2}$ hour workshops and a 2-day reflection workshop in the second year, and four 3-hour workshops and two $2\frac{1}{2}$ hour review workshops in the third year). During the workshops, the teachers were shown extracts of videotapes selected to illustrate critical aspects of children's thinking. Teachers were then prompted to reflect on what they had seen, by, for example, being challenged to relate the way a child had solved one problem to how they had solved, or might solve, other problems (Fennema, Carpenter, Franke, Levi, Jacobs & Empson, 1996 p. 407). Throughout the project, the teachers were encouraged to make use of the evidence they had collected about the achievement of their students to adjust their instruction to better meet their students' learning needs.

The teachers in the CGI program taught problem-solving significantly more and number facts significantly less than did controls. They also knew more about individual students' problem-solving processes, and their students did better in number fact knowledge, understanding, problem solving, and confidence (Carpenter, Fennema, Peterson, Chiang & Loef, 1989). More importantly, four years after the end of the program, the participating teachers were still implementing the principles of the program (Franke, Carpenter, Levi & Fennema, 2001).

Another study that showed the substantial benefits of adapting instruction to meet student learning needs was that conducted by Bergan, Sladeczek, Schwarz & Smith (1991). The performance of 428 Kindergarten students taught by 29 teachers implementing a measurement and planning system (MAPS) was compared with that of 410 students taught by 27 teachers who taught their classes as usual. In the MAPS program, teachers, together with an aide and a site manager, assessed their students' readiness for learning in reading and mathematics in the fall, and again in the spring (children were allowed as much time as they needed to complete the items). Teachers in the experimental group were trained on how to interpret the test results, and provided with the *Classroom Activity Library*—a series of activities typical of early grades instruction, but keyed specifically to empirically

validated developmental progressions—which they could use to individualize instruction, depending on the students' performance in the assessments. At the end of the year, 111 of the students in the control group (27%) were referred for placement in a special education program for the following year, and 80 (20%) were actually placed in special education programs. In the experimental group, only 25 students (6%) were referred, and only 6 students (1.4%) were placed in special education programs. In other words, students in the control group were 4.5 times more likely to be referred for placement in special education, and 14 times more likely actually to be placed in special education programs than students taught by teachers using the MAPS scheme. What is perhaps even more remarkable about this study is that all the schools in the study served districts with considerable socio-economic disadvantage, and the socioeconomic status of the students in the experimental group was actually lower than that of the control group.

A third example of the use of assessment to improve student learning was a project involving a group of 24 (later expanded to 36) secondary school mathematics and science teachers in six schools in two districts in England (Black, Harrison, Marshall, Lee & Wiliam, 2003).

The work with teachers had two main components. The first was a series of eight workshops over an 18-month period (from February 1999 to June 2000). Seven of the workshops were of 5 hours duration and one was of 3 hours duration. During the workshops, the teachers were introduced to the research basis underlying how assessment can support learning (derived from Black and Wiliam, 1998a; 1998b), had the opportunity to develop their own plans, and, at later meetings, to discuss with colleagues the changes they had attempted to make in their practice.

The second component of the intervention with the teachers was a series of visits by researchers to the teachers' classrooms, so that the teachers could be observed implementing some of the ideas they had discussed in the workshops, had an opportunity to discuss their ideas, and could plan how they could be put into practice more effectively

A key feature of the workshops was the development of action plans. As Perrenoud (1998) has pointed out, changing pedagogy requires teachers to re-negotiate the "learning contract" (cf Brousseau, 1984) that they have evolved with their students, suggesting that radical changes are best effected at the beginning of a new school year.

For the first six months of the project, therefore, the teachers were encouraged to experiment with some of the strategies and techniques suggested by the research, such as rich questioning, providing feedback to students in the form of comments rather than scores or grades, sharing learning intentions and success criteria with learners, and student peer- and self-assessment (see below). Each teacher was then asked to draw up, and later to refine, an action plan specifying which aspects of formative assessment they wished to develop in their practice and to identify a focal class with whom these strategies would be introduced in September 1999. Most of the teachers' plans contained reference to two or three important areas in their teaching where they were seeking to increase their use of formative assessment, generally followed by details of techniques that would be used to make this happen. In almost all cases the plan was given in some detail, although many teachers used phrases with meanings that differed from teacher to teacher (even within the same school).

Almost every teacher's plan contained some reference to focusing on or improving the teacher's own questioning techniques although only 11 gave details on how they were going to do this (for example using more open questions, allowing students more time to think of answers or starting the lesson with a focal question). Others were less precise (for example stating that they intended using more sustained questioning of individuals, or improving questioning techniques in general). Some teachers mentioned planning and recording their questions. Many teachers also mentioned involving students more in setting questions (for homework, or for each other in class). Some teachers also saw existing standardized tests as a source of good questions.

Nearly half the teachers mentioned providing feedback in the form of comments rather than scores or grades, although only 6 of the teachers included it as a specific element in their action plans. Some of the teachers wanted to reduce the use of scores and grades, but foresaw problems with this, given school policies on assessment. Four teachers planned for a module test to be taken before the end of the module thus providing time for remediation.

Sharing the objectives of lessons or topics was mentioned by most of the teachers, through a variety of techniques (using a question that the students should be able to answer at the end of the lesson, stating the objectives clearly at the start of the lesson, getting the students to round up the lesson with an account of what they had learned). About half the plans included references to helping the students understand the rubrics used for investigative or exploratory work, generally using exemplars from the work of students from previous years. Exemplar material was mentioned in other contexts such as having work on display and asking students to assess work using rubrics provided by the teacher.

Almost all the teachers mentioned some form of self-assessment in their plans, ranging from using red, yellow or green "traffic lights" to indicate the student's perception of the extent to which a topic or lesson had been understood, to strategies that encouraged self-assessment via targets which placed responsibility on students (e.g. "One of these twenty answers is wrong: find it and fix it!"). Traffic lights were mentioned in about half of the plans and in practically all cases their use was combined with strategies to follow up the cases where the students signaled incomplete understanding.

Several teachers mentioned their conviction that group work provided important re-inforcement for students, as well as providing the teacher with insights into their students' understanding of the work.

The other component of the intervention, the visits to the schools, provided an opportunity for researchers to discuss with the teachers what they were doing, and how this related to their efforts to put their action plans into practice. The interactions were intended to be supportive rather than directive, but since researchers were frequently seen as "experts" in either mathematics or science education, there was a tendency sometimes for teachers to invest questions from a member of the project team with a particular significance, and for this reason, these discussions were often more effective when science teachers were observed by mathematics specialists, and vice-versa.

A detailed description of the qualitative changes in teachers' practices is beyond the scope of this chapter (see Black, Harrison, Lee, Marshall & Wiliam, 2003 for a full account). In quantitative terms, students taught by the teachers developing the use of assessment for learning outscored comparable students in the same schools by approximately 0.3 standard

deviations, both on teacher-produced and external state-mandated tests (Wiliam, Lee, Harrison & Black, 2004). Since one year's growth in mathematics as measured by TIMSS is approximately one-third of a standard deviation (Rodriguez, 2004 p.18), the effect of the intervention can be seen to almost double the rate of student learning.

Formative assessment: prevalence and impact

These studies show that integrating assessment with instruction is both possible and desirable. It is also valued. A sample of 580 principles ranked "Determining what needs to be re-taught after tests" as the most important in a list of 26 assessment competences desired in their teachers (Marso & Pigge, 1993 pp. 137-138), although teachers rated competence in grading as the most important. However, there is little evidence of such data-driven practices being regularly enacted in classrooms. In a survey first published in 1980, Salmon-Cox stated that "student scores on standardized tests are not very useful to the classroom teacher" and concluded that teachers prefer to rely on their own judgment about student weaknesses and areas of needed help (Salmon-Cox, 1981 p. 631). Stiggins and Bridgeford (1985) found that while many teachers created their own assessments, "in at least a third of the structured performance assessments 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).

McMorris & Boothroyd (1993) analyzed the quality of tests developed by seventh and eighth grade mathematics and science teachers, and found that science teachers made greater use of multiple-choice items while mathematics teachers tended to set more computation items. However, for both mathematics and science teachers, the tests were of variable quality, with a significant correlation between test quality and the amount of training in educational measurement the teachers had received.

More recently Senk, Beckman and Thompson (1997) conducted a survey of assessment practices in 19 mathematics classes in 5 high schools in 3 states. They found that formal tests and quizzes were the most frequently used assessment tools, with a nominal weight across the sample of 77%, while written projects and interviews accounted for a further 7%. The tests and quizzes used focused largely on "low-level" aspects of the domains assessed, and the "grading" function of assessment dominated the "assessment" function (58% of all assessments were reported in terms of a "brute grade"). The survey also found that teachers tended to ignore the results of standardized tests in arriving at terminal grades.

In a national survey, Dorr-Bremme & Herman (1986) found that students spent around 12 hours each year taking mathematics tests in elementary school (4th to 6th grade), and approximately twice that in 10th grade, although only about 6 hours was required by the state or district (pp. 16-17), and two substantial review articles, one by Natriello (1987) and the other by Crooks (1988), provided clear evidence that classroom assessment 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 assessment 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.

Bangert-Drowns, Kulik, Kulik and Morgan (1991) reported the results of a meta-analysis of 40 research reports on the effects of feedback in what they called 'test-like' events (eg evaluation questions in programmed learning materials, review tests at the end of a block of teaching, etc). They found that providing feedback in the form of answers to the review questions was effective only when students could not look ahead to the answers before they had attempted the questions themselves (what they called "controlling for pre-search availability"). Furthermore feedback was more effective when the feedback gave details of the correct answer, rather than simply indicating whether the student's answer was correct or incorrect. In the studies where the students could not look ahead for the answers, and the feedback gave details of the correct answer, the mean effect size was 0.58 standard deviations. Reviews by Dempster (1991, 1992) confirm these findings, as does a review by Elshout-Mohr (1994) which reports many findings not available in English.

The difficulty of reviewing relevant research in this area was highlighted by Black and Wiliam (1998a), in their synthesis of research published since the reviews by Natriello and Crooks. Those earlier two 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.

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 formative classroom assessment practices can indeed have a substantial impact on student achievement.

What is less clear is what exactly constitutes *effective* classroom assessment. Although the studies cited above indicate that assessment for learning can improve learning, several studies have found conflicting results. For example, in a study of 32 fifth grade teachers in Germany, Helmke and Schrader (1987) found that teachers who had an accurate knowledge of their students (as measured by the teachers' ability to predict achievement test scores) was associated with higher levels of achievement *only* when the teachers also showed a high range of instructional techniques. Students taught by teachers who had a high knowledge of their students' achievement, but lacked a range of instructional techniques actually performed worse than students taught by teachers who did not know their students'

achievement. It appears from this study that collecting data if you can't do anything with it is counter-productive.

Furthermore, even when teachers do manage to use information about student achievement to adjust or individualize their instruction, teachers may lack the ability to do so effectively. For example, in a 20-week study of 33 teachers in elementary and middle schools, Fuchs, Fuchs, Hamlett and Stecker (1991) found that teachers who received feedback on the achievement of students in their classes with learning difficulties made more adjustments to their teaching programs than teachers not given this information. However, the achievement of these students was improved *only* when this feedback was accompanied by advice from a computerized "expert system", since the teachers not given the feedback from the expert system tended to re-explain how to do problems with the same algorithms that had led to previous failure.

Therefore, if we are to maximize the potential benefits of formative assessment or assessment for learning, there is a need to understand what, exactly, constitutes effective formative assessment, and this is the focus of the remainder of this chapter.

Formative assessment: theoretical considerations

In education, the term "feedback" is routinely applied to any information that a student is given about their performance, and the generality of the term obscures that this is, in reality a frozen metaphor, derived from systems engineering. One of the earliest uses of the term was by Norbert Wiener. In 1940, Wiener and his colleagues had been working on automatic range-finders for anti-aircraft guns, which involved mechanisms for predicting the path of airplanes, and realized that the control mechanisms needed for the range-finders were similar to control mechanisms in animals. He realized that purposeful action required the existence of "a closed loop allowing the evaluation of the effects of one's actions and the adaptation of future conduct based on past performances" (Wiener, 1948). He also realized that there were two kinds of loops: those involving positive feedback and those involving negative feedback. In both kinds of loop, there are inputs to the system, and, some time later, outputs from the system. The crucial feature of the loop is that information about the output is fed back to the input side of the system. In positive feedback loops, the feedback serves to drive the system further in the direction it is already going, while in negative feedback loops, the feedback acts to oppose the current direction of the system. All positive feedback systems are unstable, driving the system either towards explosion or collapse. Examples of the explosive kind of positive feedback loops are simple population growth in the presence of plentiful supplies of food and in the absence of predators, and inflationary price/wage spirals in economics. Examples of collapse are economic depression, food hoarding in times of shortage, and the loss of tax revenue in urban areas as a result of "middle-class flight" (note that whether the effects are explosion or collapse, both are examples of positive feedback).

In contrast, negative feedback systems tend to produce stability, because they are inherently "self-correcting". One example of negative feedback is population growth with limited food supply, in which the lack of food causes a slow-down in population growth, which in turn, depending on the conditions, produces either an asymptotic approach towards, or a damped oscillation about, the "carrying capacity" of the environment. Another example is the domestic thermostat. When the temperature of the room drops below the setting on the

thermostat, a signal is sent to turn on the furnace. When the room heats up above the setting on the thermostat, a signal is sent to turn off the furnace.

From the foregoing discussion, it will be clear that the current uses of the term "feedback" in education are very different from those in engineering, but more importantly, the simplistic engineering metaphor may just not be helpful. In systems engineering, negative feedback is good, because it keeps a system under control while positive feedback is bad because it leads to explosion or collapse. In educational settings, things are not so clear-cut. Negative feedback may be helpful for correcting learning when it is off-course, but feedback that reinforces learning that is on track is also powerful. And the thermostat doesn't care how often it is told that the temperature in the room is "wrong" and needs to be corrected, whereas we know that humans are often adversely affected by such information. This is not to say that the insights of systems engineering are irrelevant, but we do need to exercise considerable caution in adopting what are essentially metaphors in complex areas like human learning, especially in terms of the relationship between "feedback" and "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 termed "knowledge of results"). These kinds of usages suggest that the distinction between "formative" and "summative" applies to the assessments themselves, but since the same assessment can be used both formatively and summatively, these terms are more usefully applied to the *use* to which the information arising from assessments is put.

As Ramaprasad (1983) notes, the defining feature of feedback is that the information generated within the system must have some effect. Information that does not have the capability to change the performance of the system is 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).

Commenting on this, Sadler (1989) noted:

> An important feature of Ramaprasad's definition is that information about the gap between actual and reference levels is considered as feedback *only when it is used to alter the gap*. If the information is simply recorded, passed to a third party who lacks either the knowledge or the power to change the outcome, or is too deeply coded (for example, as a summary grade given by the teacher) to lead to appropriate action, the control loop cannot be closed, and "dangling data" substituted for effective feedback (p121, emphasis in original).

In this view, formative assessments (or 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 and Black, 1996). Therefore what

is important is not the intent behind the assessment, but the function it actually serves, and this provides a useful point of difference between the terms "assessment for learning" and "formative assessment". Black, Harrison, Marshall, Lee 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 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. that 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.

Another way of thinking about the distinction being made here is in terms of monitoring assessments, diagnostic assessments and formative assessments. 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. However, if a pitching coach is able to see that she is not dropping her 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*. If the athlete is able to use the advice about the delivering the pitch from below the knee to make her rising fast-ball rise, then the

feedback given by the coach will, indeed, have been formative. This use of formative recalls the original meaning of the term. In the same way that one's formative experiences are the experiences that shape the individual, formative assessments are those that shape learning.

The important point here 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. If we want to promote learning, we have to collect the correct data in the first place. This is discussed further in the section on *Clarifying and sharing learning intentions* below.

In the examples given above the action follows quite quickly on from the elicitation of the evidence about student achievement, but the definition of formative assessment given above allows cycles of elicitation, interpretation and action of any length, provided the information is used to form the direction of future learning.

For example, in March, a mathematics supervisor may be planning the workshops she will make available to teachers in the summer. By looking at the scores obtained by the students in the district on last year's state-mandated tests, she might discover that, compared to other students in the state, the students are performing relatively poorly on items that assess geometry. As a result, she might plan a series of workshops on teaching geometry for the summer, and, if these workshops are successful, then the students' performance on geometry items should improve. This cycle would be over two years in length, since the supervisor would be using data from tests taken the previous March, and the impact would not be felt until the students took the test the following March (and the results might not be available until July). Furthermore, the data that was used to improve learning in the district were not collected from the students who benefited, but those who were in that grade two years earlier. Nevertheless, data from the state tests functioned formatively in this example, since information about student performance was used to make adjustments to instruction (in this case improving the teaching of geometry) that improved the learning of mathematics in the district. It could be argued that this kind of usage is *evaluation*, rather than *assessment*, since it is the program that is being improved through the analysis of student data, but there is no clear boundary between the two. A teacher might look through the responses of her students 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 could still 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. At this level, both the program and the learning of individual students on whom the data were collected is being impacted by the assessment outcomes, so this example could be thought of as either assessment or evaluation. The crucial point is that the evidence is used to make decisions that could not be made, or at least could not be made as well, without that evidence, with the result that learning is enhanced.

The building-in of time to make use of assessment data 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.

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, but 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 1.

Type	Focus	Length
Long-cycle	across instructional units, quarters, semesters, years	four weeks to one year
Medium-cycle	between lessons or units	one day to four weeks
Short-cycle	within a single lesson	five seconds to two hours

Table 1: Cycle lengths for formative assessment

The foregoing discussion establishes that any assessment can be formative, and that an assessment is formative to the extent that information from the assessment is used to adjust instruction to better meet student learning needs. The adjustment can take place immediately, in time for the next instructional episode, between units, or even between years, and the beneficiaries of the adjustments may be the students on whom information was collected, or they may not. However, while all these different uses of assessment information may be formative, in that the information is used to adapt instruction to better meet student learning needs, the research evidence cited above suggests that not all are equally effective. To establish what, exactly, constitutes the most effective form of formative assessment requires a deeper look at the research evidence.

An important feature of Ramaprasad's definition of feedback (what we are calling here "formative assessment") 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 primarily the teacher's role, but we need also to take account of the role that the learners themselves, and their peers, play in these processes. Crossing the three key instructional process listed above with the three agents involved in the classroom produces the framework shown in figure 1, which provides a way of thinking about the key strategies involved in formative assessment. The subject classroom that is the focus of figure 1 is, of course, itself nested within a school, which in turn is located in a community, and so on. Any adequate account of formative assessment will have to acknowledge these multiple contexts, but they are beyond the scope of this chapter. Furthermore, since 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, at least as a first order approximation. For examples of sociocultural approaches to the implementation of formative assessment, see Black and Wiliam (2005b) and Pryor and Crossouard (2005).

This framework suggests that assessment for learning can be conceptualized as consisting of five key strategies and one "big idea" (Wiliam & Thompson, to appear). The five key strategies are

1. engineering effective classroom discussions, questions, and learning tasks that elicit evidence of learning;
2. providing feedback that moves learners forward;
3. clarifying and sharing learning intentions and criteria for success;
4. activating students as the owners of their own learning; and
5. activating students as instructional resources for one another.

<table>
<tr><th></th><th>Where the learner is going</th><th>Where the learner is right now</th><th>How to get there</th></tr>
<tr><td>Teacher</td><td>Clarifying learning intentions and sharing and criteria for success</td><td>Engineering effective classroom discussions and tasks that elicit evidence of learning</td><td>Providing feedback that moves learners forward</td></tr>
<tr><td>Peer</td><td rowspan="2">Understanding and sharing learning intentions and criteria for success</td><td colspan="2" rowspan="2">Activating students as instructional resources for one another</td></tr>
<tr><td>Learner</td></tr>
<tr><td></td><td>Understanding learning intentions and criteria for success</td><td colspan="2">Activating students as the owners of their own learning</td></tr>
</table>

Figure 1: Aspects of assessment for 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. The following sections describe in more detail these five key strategies and how they fit together within the more general idea of the "regulation of learning processes."

Engineering effective classroom discussions and tasks that elicit evidence of learning

There is a vast range of strategies that teachers can use to elicit evidence of student learning, from formal testing occasions, through the activities that students routinely undertake in mathematics classrooms, to classroom discussions and informal exchanges with students. In the previous version of this handbook, Webb (1992) outlined a set of principles for assessing mathematics. The chapters by Jan de Lange and Linda Dager Wilson in this volume cover large-scale and high-stakes assessment respectively and Clarke (1996) provides an international perspective on contemporary mathematics assessment. In addition, van den Heuvel-Panhuizen & Becker (2003) make proposals for a didactic model of assessment design in mathematics and an excellent overview of the characteristics of, and general principles for the design of "thought-revealing activities" can be found in Lesh, Hoover, Hole, Kelly & Post (2001). The focus of this section is on classroom discussions and tasks that elicit evidence of learning.

Classroom discussions

Teacher-led classroom discussion, along with individual seat-work, is one of the staples of mathematics instruction in the USA (and indeed most other countries). In the traditional model of classroom transactions—termed Initiation-Response-Evaluation or I-R-E by Mehan (1979)—the teacher asks a question, chooses a student to answer the question, and then makes some response to the student's answer. Within this broad structure we can say rather more.

First, it is important to note that the teacher almost invariably dominates such classroom discussions, even in the US, where it has been shown that American teachers actually talk less than teachers in countries that are more successful in mathematics, at least as measured in international comparisons. For example, the 1999 TIMSS video study found that in US classrooms there were 8 teacher words for every student word, while in Japan, there were 13, and in Hong Kong 16 teacher words for every student word (Hiebert, Gallimore, Garnier, Givvin, Hollingsworth, Jacobs, Chui, Wearne, Smith, Kersting, Manaster, Tseng, Etterbeek, Manaster, Gonzales, & Stigler, 2003). Of course, this means that even in a US classroom of 25 students, the teacher speaks 200 times as much as any one student. It also makes clear that the quality of what the teacher is saying is much more important than the quantity, and here, the evidence from the TIMSS video studies are compelling.

Although the 1999 TIMSS video study focused on eighth grade, the findings were similar to those of earlier studies (e.g. Weiss, Pasley Smith, Banilower & Heck, 2003; Rowan, Harrison & Hayes, 2004), and showed that mathematics teaching in the US "is characterized by frequent review of relatively unchallenging, procedurally oriented mathematics" (Hiebert, Stigler, Jacobs, Givvin, Garnier, Smith, Hollingsworth, Manaster, Wearne & Gallimore, 2005 p. 125). Why this should be the case is not clear, but the consequences are profound. To see why, it is necessary to discuss briefly some recent work on the nature and development of human abilities.

The term "intelligence" has been used in a variety of ways for at least a hundred years, and there is no consensus on its meaning today. The word has been tarnished by its use by the eugenics movement in the first half of the twentieth century, especially in the United States (see Selden, 1999 for an extended discussion), and more recently, the term has been used in

essentially racist projects such as "The Bell Curve" by Herrnstein & Murray (1994). There is ample evidence that the conclusions reached by such authors is fundamentally flawed (see, for example, Montague, 1999; Fish, 2001) but the intensity and the politicization of the debate makes it difficult to separate the science from the myth. One particularly unfortunate effect of this debate is that it has led many people to adopt rigid and extreme positions on the issue of intelligence.

At one extreme are those who believe that:

- Intelligence is determined entirely by one's genes, and is fixed for life;
- Intelligence tests measure the most important aspects of human thinking;
- Intelligence is the most important predictor of success in the workplace; and
- Other kinds of ability don't really matter.

At the other extreme, partly in response to the political motives of those who hold the extreme views listed above, are those who reject the concept of intelligence entirely, or deny its relevance except in the most limited laboratory studies. For these proponents:

- Intelligence is determined by the environment, and not one's genes;
- Intelligence tests measure only the ability to take intelligence tests;
- Intelligence doesn't matter in the real world; and
- There are several different kinds of intelligence, all independent of each other

It turns out that there ***is*** a high degree of consensus amongst psychologists on the science underlying intelligence, and, predictably, it is somewhere between the two extreme positions described above. It is that:

- Intelligence is determined by both environment and genetics, and the genetic influence is substantial (Plomin & Petrill, 1997);
- Intelligence tests correlate strongly with a range of other measurements of mental capability (Mackintosh, 2000);
- Intelligence is strongly associated with success in a wide range of real world activities (Bertua, Anderson & Salgado, 2006);
- There are several different aspects of intelligence, but most of them are strongly inter-related (Deary, 2000).

In this sense, intelligence appears to be like physical height. Physical height is inherited—tall parents tend to have tall children—but there is also a strong environmental component. The better the standard of nutrition, the taller the child grows. The same is clearly true for intelligence, most clearly evidenced by the "Flynn effect". Named after James Flynn, this is the name given to the rise of about one standard deviation that has been observed in IQ scores throughout the developed world over the last sixty years (Flynn 1984; 1987). While there is disagreement about the causes of these gains (see, for example, Neisser, 1998), it is clear that students' educational experiences are an important element.

Dickens and Flynn (2001) have shown that what we observe about intelligence is best accounted for by the idea that people select, or have selected for them, environments that match their intelligence. People with high intelligence, for example, engage in more of the activities that enhance intelligence, and so become more intelligent whereas people with lower intelligence opt out of, or are denied, these intelligence-enhancing activities and so

lose the opportunities to enhance their intelligence. This suggests that intelligence and environment are mutually constitutive of each other: environment causes intelligence and intelligence causes environment. However, the model proposed by Dickens and Flynn also suggests that the impact of transient improvements in environment are themselves transient This explains why compensatory preschool programs have significant effects on intelligence while students are in the program, but the effects diminish when students leave (Barnett & Camilli, 2001), although it should be noted that the improvements in student *achievement* produced by such programs are lasting.

One concrete demonstration on the power of classroom environments to improve student achievement is a study of 191 students in seven fourth grade classrooms following the "Thinking Together" program (Dawes, Mercer & Wegerif, 2000) as they learned science. These students outperformed controls in similar schools by 0.74 standard deviations on tests of concept mapping and 0.29 standard deviations on a standardized science achievement test (Mercer, Dawes, Wegerif & Sams, 2004). Perhaps even more surprisingly, these students outperformed controls by 0.27 standard deviations on a purely spatial intelligence test—Raven's Progressive Matrices (Raven, 1960). Just by increasing the amount of structured talk in the classroom, students got better at purely spatial tasks.

The implications of these findings for mathematics classrooms are profound. If we are to maximize mathematics achievement, then classrooms must be places where every student is required to engage cognitively with the mathematics she or he is learning, and this appears to be the hallmark of practice both in countries that are successful, and exemplary classrooms in the USA (see, for example, Boaler & Humphreys, 2005). Shulman (2005) has described such cognitively rich pedagogies as *pedagogies of engagement*.

The problem is that, as is clear from the TIMSS video studies (Hiebert, Gallimore, Garnier, Givvin, Hollingsworth, Jacobs, Miu-Ying Chui, Wearne, Smith, Kersting, Manaster, Tseng, Etterbeek, Manaster, Gonzales & Stigler, 2003), few mathematics classrooms in the USA rigorously employ pedagogies of engagement. In most mathematics classrooms in the USA, the choice about whether to engage is left to the student. In classroom dialogue, the teacher asks a question, and then selects a respondent from those who have signaled that they have an answer by raising their hands. Some teachers do try to counter this by occasionally calling on students who have not raised their hands, but this is frequently seen as breaching the terms of the "didactical contract" (Brousseau, 1984). One teacher summed up his predicament thus:

> I'd become dissatisfied with the closed Q&A style that my unthinking teaching had fallen into, and I would frequently be lazy in my acceptance of right answers and sometimes even tacit complicity with a class to make sure none of us had to work too hard ... They and I knew that if the Q&A wasn't going smoothly, I'd change the question, answer it myself or only seek answers from the "brighter students". There must have been times (still are?) where an outside observer would see my lessons as a small discussion group surrounded by many sleepy onlookers. (Black, Harrison, Lee, Marshall & Wiliam, 2004 p. 11)

The consequence of this is that some students are deeply engaged in the lesson, and as such, are increasing their capabilities, while others are avoiding engagement, and thus forgoing the opportunities to increase their ability.

In other classrooms, participation is not voluntary. Magdalene Lampert (2001) describes a lesson she taught in which she made a point of calling on a student who had not raised his

hand even though many others had done so. She gave her reasons as follows: "I called on Richard because I wanted to teach him and others in the class that everyone would indeed be asked to explain their thinking publicly. I also wanted to teach everyone that what they said would be expected to be an effort to make mathematical sense" (p. 146).

Leahy, Lyon, Thompson & Wiliam (2005) describe mathematics classrooms in which teachers have gone further, and instituted a rule of "no hands up, except to *ask* a question." After posing a question, the teacher decides which student should respond by the use of some randomizing device, such as name cards (see Webb, 2004 p. 175) or a beaker of Popsicle sticks on which the students' names are written. The important point about such classrooms is that mental participation is not optional.

Such a radical change in the "classroom contract" (Brousseau, 1984) may be unwelcome for many students, used to classrooms where participation is optional, but there is evidence that students' participation practices in mathematics classrooms are malleable (Turner & Patrick, 2004). In particular, there are many strategies that teachers can use to engage students in classroom participation. Where students reply "I don't know", Ellin Keene suggests that teachers can ask, "OK, but if you did know what would you say?" (Carol, 2006), or the teacher can solicit answers from other students and then return to the original student and ask them to select from amongst the answers they have heard. Other possibilities are allowing students to "phone a friend", or, for multiple-choice items, they can "ask the audience" or ask to go "fifty-fifty" where two incorrect responses are removed. All these strategies derive their power from the fact that classroom participation is not optional, and even when the student resists, the teacher looks for ways to maintain the student's engagement.

How much time a teacher allows a student to respond before evaluating the response is also important. It is well known that teachers do not allow students much time to answer questions (Rowe, 1974), and, if they don't receive a response quickly, they will "help" the student by providing a clue or weakening the question in some way, or even moving on to another student. However, what is not widely appreciated is that the amount of time between the student providing an answer and the teacher's evaluation of that answer is just as, if not more, important. Of course, where the question is a simple matter of factual recall, then allowing a student time to reflect and expand upon the answer is unlikely to help much. But where the question requires thought, then increasing the time between the end of the student's answer and the teacher's evaluation from the average "wait-time" of less than a second to three seconds, produces measurable increases in learning, although according to Tobin (1987) increases beyond three seconds have little effect, and may cause lessons to lose pace.

In fact, questions need not always come from the teacher. There is substantial evidence that getting students to generate their own questions enhances their learning. Rosenshine, Meister & Chapman (1996) found that training students to generate questions while reading increased performance by 0.36 standard deviations on standardized tests, and by 0.86 standard deviations on tests developed by the experimenters. Foos, Mora & Tkacz (1994) compared four strategies for helping students prepare for a test or examination:

a) Tell the students to review the material on which the test is based
b) Provide the students with study methods and materials

c) Tell the students to generate their own study outlines, and
d) Tell the students to generate their own study questions, with answers

They found that strategy (d) generated the highest performance, followed, in turn, by (c), (b), and (a).

Such student-produced assessments can also be useful to the teacher, because they provide useful information about what the students think they have been learning, which may not be the same as what the teacher thinks the students have been learning.

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

Another key feature of classroom questioning is the way that teachers listen to student answers. As many authors have pointed out, when teachers listen to student responses, they attend more to the correctness of the answers rather than what they can learn about the student's understanding (Even & Tirosh, 1995; 2002; Heid, Blume, Zbiek & Edwards, 1999). In a detailed study, Davis (1997) followed the changes in the practice of one middle school mathematics teacher, focusing in particular on how the teacher responded to student answers. Initially, the teacher's reactions tended to focus on the extent to which the student responses accorded with the teacher's expectations. After sustained reflection and discussion with the researcher over a period of several months, the teacher's reaction placed increasing emphasis on "information-seeking" as opposed to the "response-seeking" that characterized the earlier lessons. Davis termed these two kinds of listening "evaluative listening" and "interpretive listening" respectively. Towards the end of the two-year period, there was a further shift in the teacher's practice, with a marked move away from clear lesson structures and pre-specified learning outcomes, and towards the exploration of potentially rich mathematical situations, in which the teacher is a co-participant. Most notably, in this third phase, the teacher's own views of the subject matter being "taught" developed and altered along with that of the students (what Davis termed "hermeneutic listening"). Similar trajectories of change from evaluative to interpretive listening have been observed in other mathematics teachers (English & Doerr, 2004), in pre-service teachers (Crespo, 2000), and were eloquently summarized by a girl in the seventh grade: "When Miss used to ask a question, she used to be interested in the right answer. Now she's interested in what we think" (Hodgen & Wiliam, 2006 p. 16).

What makes a good question?

Careful analysis of students' incorrect responses to standard classroom items can reveal important insights into students' conceptions. For example, DeCorte & Verschaffel (2006) point out when students are asked to respond to the following item:

$$\_\_\_\_ - 12 = 7$$

many students write 18 or 5 in the blank space. Both responses are, of course, incorrect, but in completed different ways, the first probably indicating an arithmetical slip, and the second more likely indicating a lack of understanding of the meaning of the equal sign.

However, understanding the thinking behind students' responses is more straightforward when the questions we ask students have been planned carefully ahead of time expressly for this purpose. Two items used in the Third International Mathematics and Science Study (TIMSS), shown in figure 2 below, illustrate one important feature of the design of good questions . Although apparently quite similar, the success rates on the two items were very different. For example, in Israel, 88% of the students answered the first items correctly, while only 46% answered the second correctly, with 39% choosing response (b) (Vinner, 1997). Vinner suggests that the reason for this is that many students, in learning about fractions, develop the naive conception that the largest fraction is the one with the smallest denominator, and the smallest fraction is the one with the largest denominator. This approach leads to the correct answer for the first item, but leads to an incorrect response to the second. Further evidence for this interpretation is given by noting that 46% plus 39% is very close to 88%, suggesting that almost half of the students who answered the first item correctly did so with an incorrect strategy. In this sense, the first item is a much weaker item than the second, because students can get it right for the wrong reasons.

Item 1 (success rate 88%)

Which fraction is the smallest?

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

Item 2 (success rate 46%)

Which fraction is the largest?

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

Figure 2: two items from the Third International Mathematics and Science Study

This illustrates a very general principle in teachers' classroom questioning. By asking questions of students, teachers try to establish whether students have understood what they are meant to be learning. In other words, the teacher is trying to construct a model of the student's thinking. As von Glasersfeld (1987) notes:

> Inevitably, that model will be constructed, not out of the child's conceptual elements, but out of the conceptual elements that are the interviewer's own. It is in this context that the epistemological principle of *fit*, rather than *match* is of crucial importance. Just as cognitive organisms can never compare their conceptual organisations of experience with the structure of an independent objective reality, so the interviewer, experimenter, or teacher can never compare the model he or she has constructed if a child's conceptualisations with what actually goes on in the child's head. In the one case as in the other, the best that can be achieved is a model that remains viable within the range of available experience. (p13)

The key phrase here is "within the range of available experience." If the teacher asks questions that are more like the first TIMSS item above than the second, the range of available experience is narrow, and range of models that fit, correspondingly large. As Gay

& Thomas (1993) pointedly ask, "Just because they got it right, does that mean they know it?" (p. 130).

For teacher questioning to be effective, teachers need to know what kinds of conceptualizations students are likely to have, and need tools to identify them.

Consider, for example, the following pair of simultaneous equations:

$$3a = 24$$

$$a + b = 16$$

Many students find this difficult, saying that it cannot be done. The teacher might conclude that they need some more help with equations of this sort, but with this particular pair of equations, a more likely reason for the difficulty is not with mathematical skills but with the student's *beliefs* (see also Schoenfeld, 1985) If the students are encouraged to talk about their difficulty, they often say things like, "I keep on getting b is 8, but it can't be because a is." The reason that many students have developed such a belief is, of course, that before they were introduced to solving equations, they had been practicing substitution of numbers into algebraic formulas, where each letter always did stand for a different number. Although the students will not have been taught that each letter must stand for a different number, they have generalized implicit rules from their previous experience, just as because we always show them triangles where the lowest side is horizontal, they talk of "upside-down triangles" (Askew and Wiliam, 1995).

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

Some people have argued that these unintended conceptions are the result of poor teaching. If only the teacher had phrased their explanation more carefully, had ensured that no unintended features were learned alongside the intended features, then these misconceptions would not arise, but this argument fails to acknowledge two important points. The first is that this kind of over-generalization is a fundamental feature of human thinking. When young children say things like "I spended all my money," they are demonstrating a remarkable feat of generalization. From the huge messiness of the language that they hear around them, they have learned that to create the past tense of a verb, one adds "d" or "ed." In the same way, if one asks young children what causes the wind, a common answer is "trees." They have not been taught this, but have observed that trees are swaying when the wind is blowing and have inferred (incorrectly in this case) a causal relationship from a correlation.

The second point is that even if we wanted to, we are unable to control the student's environment to the extent necessary for unintended conceptions not to arise. For example, it is well known that many students believe that the result of multiplying 2.3 by 10 is 2.30. It is highly unlikely that they have been taught this. Rather this belief arises as a result of observing regularities in what they see around them. The result of multiplying whole-numbers by 10 is just to add a zero, so why shouldn't that work for all numbers? The only

way to prevent students from acquiring this "misconception" would be to introduce decimals before one introduces multiplying single-digit numbers by 10, which is clearly absurd. The important point is that we must acknowledge that what students learn is not necessarily what the teacher intended, and it is essential that teachers explore students' thinking before assuming that students have "understood" something.

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

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

In generating good questions, the traditional concerns of reliability and validity do not provide sound guidance as to what makes a good question. For example, many teachers think that the following question, taken from the Chelsea Diagnostic Test for Algebra (Hart, Brown Kerslake, Küchemann & Ruddock, 1985), is "unfair":

Simplify (if possible): 2a + 5b

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

Asking students which of the following two fractions is the larger raises similar issues:

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

In some senses this is a "trick question." There is no doubt that this is a very hard item, with typically only around one 14-year old in six able to give the correct answer, compared with around three-quarters of 14-year-olds being able to select correctly the larger of two "typical" fractions—i.e where the numerators and denominators are different, and each less than 12 (Hart, 1981). It may not, therefore, be a very good item to use in a test of students' achievement. But it is very important for the teacher to know if her students think that $\frac{3}{11}$ is larger than $\frac{3}{7}$. The fact that this item is seen as a "trick question" shows how deeply the summative function of assessment is ingrained into the practice of most teachers.

A third example, which caused considerable disquiet amongst teachers when it was used in a national test in England in the 1990s, is based on the following item, again taken from one of the Chelsea Diagnostic Tests:

Which of the following statements is true:

(1) AB is longer than CD

(2) AB is shorter than CD

(3) AB and CD are the same length

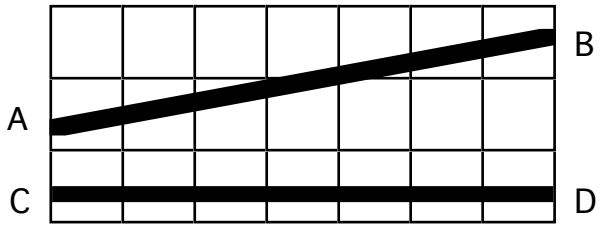

Figure 3: Item adapted from the CSMS tests (Hart et al, 1985)

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

All of the questions discussed so far have the potential to elicit evidence of active student conceptions that may hinder learning, and as such, have the potential to support teachers' instructional decision-making. However, the questions discussed above, and most of the questions used by teachers, are really only valuable when the teacher can ask the student to explain or elaborate their answers. Such questions are therefore good as *discussion questions* but less good as *diagnostic questions* (Ciofalo & Wylie, 2006). If it is necessary for each student to explain the reason for her or his answer, then it will take a great deal of time to get around the whole class. Furthermore, those students answering last are more likely to be choosing their justifications from amongst those already given by their peers than thinking hard about their own answers.

Consider the item shown in figure 4. This item has some merit as a discussion item, since it is likely to elicit very quickly the fact that some students in a class think that $a^2 + b^2 = c^2$ for this triangle because it is true for all right triangles. However, as soon as students hear a peer mention that $a^2 + b^2 = c^2$ is true only when *c* is the hypotenuse, then they may well give the same response.

In the right triangle below, does $a^2 + b^2 = c^2$?

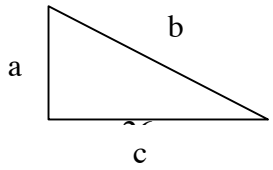

Figure 4: Item on Pythagoras' rule

As an alternative, consider the item shown in figure 5. For this item, there are 64 possible responses (the statement might be true for each of the six triangles giving a total of 2^6 possibilities), so the chance of a student getting this correct by guessing is less than 2%. Moreover, the item can be used to get a response from every single student in the class at the same time, by giving students a set of six cards labeled A, B, C, D, E, and F. This use of questions with an "all-student response system" (Leahy et al, 2005) provides teachers with much richer data on the level of understanding in a class than is possible with a "single-student response system" such as asking students to raise their hands, or selecting a student at random to answer.

In which of these right triangles is $a^2 + b^2 = c^2$?

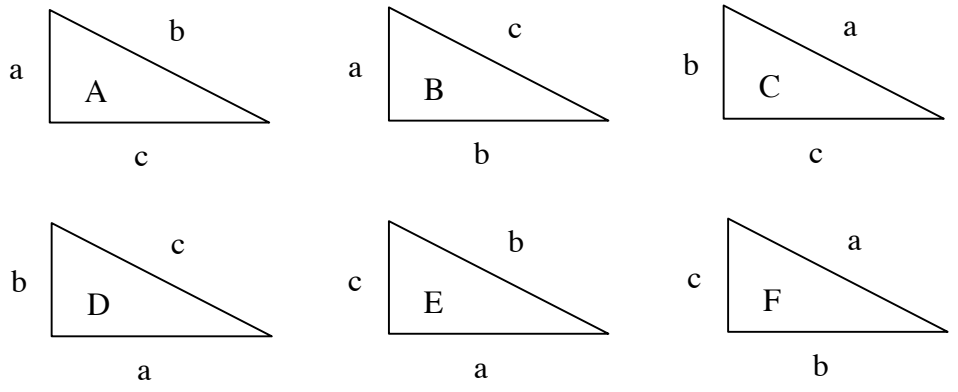

Figure 5: Item on Pythagoras' rule for use as an "all-student response" item

By moving from single-student response systems to all-student response systems, the teacher is able to make more effective instructional decisions in real-time. If all students answer correctly, then she can move on. If not, then the teacher has created a "teachable moment". If no-one answered correctly, then the teacher might choose to re-teach the material to the whole class using a different approach from the one that she used originally. However, if some of the students answer correctly, and some answer incorrectly, this provides an opportunity for a class discussion of the issue. Not only does the teacher know that some students haven't understood, but she is able to use the information about the responses of each student to engineer a more effective discussion, by, for example, calling on all those with a particular response to agree their reasons for that response before considering other responses. Of course the aggregation of information from student responses is greatly enhanced by the use of electronic "clickers" (Roschelle, Abrahamson and Penuel, 2004) but it is worth noting that most such systems can currently only cope with a single response, so that items such as that in figure 5 would be much more difficult to use. Incorrect responses to these items may reveal "facets" of student thinking (diSessa & Minstrell, 1998), and thus provide important evidence about student learning. In this sense, it is useful if the incorrect responses made by students are interpretable. However, if we accept that in teaching, assuming that students do understand something when they

don't is more damaging than assuming they don't understand something when they do, then a more important requirement is that that the *correct* response is interpretable. In other words, the teacher needs to be sure that the student got the answer right for the right reasons.

Sets of lettered cards work well with multiple-choice items, but not all items can usefully be presented in such a format. For this reason, many teachers have adopted the use of "slates" or "mini-white boards" on which students can write and display an answer for the teacher to see. Such a technique is particularly useful where students' responses are difficult to predict, such as asking the students to write down a fraction between $\frac{1}{6}$ and $\frac{1}{7}$ if they can. Some students will say that it can't be done, others will generate incorrect solutions, and others will generate correct solutions. The powerful feature of such unstructured response systems is that it allows students to generate substantially correct solutions in non-standard form, such as

$$\frac{1}{6\frac{1}{2}}$$

Perhaps the most important point in all this is that questions worth asking are unlikely to be generated spontaneously in the middle of a lesson. They need careful planning, preferably in collaboration with other teachers, as is practiced in Japanese "lesson study" (Fernandez & Makoto, 2004).

Providing feedback that moves learners forward

From the reviews of research conducted by Natriello (1987), Crooks (1988), Bangert-Drowns *et al*. (1991), and Black and Wiliam (1998a) cited above it is clear that not all kinds of feedback to students about their work are equally effective. For example, Meisels, Atkins-Burnett, Xue, Bickel & Son (2003) explored the impact of the Work Sample System (WSS)—a system of curriculum-embedded performance assessments—on the achievement of 96 third grade urban students in reading and mathematics, as measured by the Iowa Test of Basic Skills (ITBS). When compared with a sample of 116 third graders in matched schools, and with students in the remainder of the school district (Pittsburgh, PA), the achievement of WSS students was significantly and substantially higher in reading, but in mathematics, the differences were much smaller, and failed to reach statistical significance. The details of the system in use, how it is implemented, and the nature of the feedback provided to students appear to be crucial variables, with small changes often producing large impacts on the effectiveness of the system. Some insight into the characteristics of effect feedback is provided by a study by Nyquist (2003).

The reviews by Natriello, Crooks, and Black and Wiliam, had focused on K-12 education, but Nyquist found that the findings about the importance of assessment for learning generalize to higher education too. In reviewing the research on the effects of feedback in 185 studies in higher education, he developed the following typology of different kinds of formative assessment:

> *Weaker feedback only*: students are given only the knowledge of their own score or grade, often described as "knowledge of results."

Feedback only: students are given their own score or grade, together with either clear goals to work towards, or feedback on the correct answers to the questions they attempt, often described as "knowledge of correct results."

Weak formative assessment: students are given information about the correct results, together with some explanation.

Moderate formative assessment: students are given information about the correct results, some explanation, and some specific suggestions for improvement

Strong formative assessment: students are given information about the correct results, some explanation, and specific activities to undertake in order to improve.

He then calculated the average standardized effect size for the studies for each type of intervention, and the results are given in table 2 below.

	N	Effect
Weaker feedback only	31	0.16
Feedback only	48	0.23
Weaker formative assessment	49	0.30
Moderate formative assessment	41	0.33
Strong formative assessment	16	0.51
Total	185	

Table 2: standardized effect sizes for different kinds of feedback interventions

Nyquist's results echo the findings of Bangert-Drowns *et al*. discussed above. Just giving students feedback about current achievement produces very little benefit, but where feedback engages students in mindful activity, the effects on learning can be profound.

In one study (Elawar & Corno, 1985), 18 sixth grade teachers in three schools in Venezuela received seven hours of training on how to give constructive written feedback on the mathematics homework produced by their students (specific comments on errors, suggestions about how to improve and at least one positive remark). Another group of teachers graded homework as normal (i.e. just scores) and a third group gave constructive feedback to half their classes and just scores to the other half. The students receiving the constructive feedback learned twice as fast as the control group students (in other words, they learned in one week what the others would have taken two weeks to learn). Moreover, in classes receiving the constructive feedback, the achievement gap between male and female students was reduced, and attitudes towards mathematics were more positive.

The negative impact of grades was also established in a study of Israeli students conducted by Butler & Nisan (1986). In this study, one group of students received non-threatening task-related evaluations of their work, one group received normative grades, and one group received no feedback. The level of intrinsic motivation of students given grades was lower than those given the task-related evaluations, and this was also reflected in the level of achievement. Interestingly, those given no feedback also showed lower levels of intrinsic motivation—this suggests that task-related feedback has an important role to play in building intrinsic motivation (see section on *Activating*

students as owners of their own learning below) and that lack of feedback can, under some circumstances, be as deleterious as the wrong sort of feedback, but the relationship is clearly complex, since other studies have not found lack of feedback to be consistently associated with lower levels of learning. For example, Grolnick & Ryan (1987) reported a study of 91 fifth-grade children that assessed the effects of different conditions on emotional experience and performance on a learning task. Two directed-learning conditions, one controlling (DLC) and one non-controlling (DLN), were contrasted with each other and with a third non-directed, spontaneous-learning context (NLC). In both the DLC and DLN conditions, what was to be learned was specified and students knew that there would be an assessment, but in the DLC condition, students were told they would be given grades as they worked, while those in the DLN condition were told there would be no grades and that the purpose of the activity was to see what they could learn. In the NLC condition, students were exposed to material without specifications and had no knowledge of the subsequent assessment. On measures of rote learning, both the DLC and DLN groups scored higher than the NLC, but the level of interest and the amount of conceptual learning was lower in the DLC group than in the other two. Furthermore, children in the DLC condition evidenced a greater deterioration in rote learning in a follow-up assessment approximately one week later.

In a subsequent study, Butler (1988) investigated the effectiveness of different kinds of feedback in 12 sixth grade classes in four Israeli schools. For the first lesson, the students in each class were given a booklet containing a range of divergent thinking tasks. At the end of the lesson, their work was collected in. Independent scorers then assessed this work. At the beginning of the next lesson, two days later, the students were given feedback on the work they had done in the first lesson. In four of the classes students were given scores (which were scaled so as to range from 40 to 99) while in another four of the classes, students were given comments, such as "You thought of quite a few interesting ideas; maybe you could think of more ideas." In the other four classes, both scores and comments were written in the students' notebooks. Then, the students were asked to attempt some similar tasks, and told that they would get the same sort of feedback as they had received for the first lesson's work. Again, the work was collected in and assessed.

In order to explore the differences in impact of these treatments on students, the achievement and attitudes toward the subject under study of the highest-performing students in each class were compared with the lowest achievers in each class. Those given only scores made no gain from the first lesson to the second. Those who had received high scores in the tests were interested in the work, but those who had received low scores were not. The students given only comments scored, on average, 30% higher on the work done in the second lesson than on the first, and the interest of all the students in the work was high. However, those given both scores and comments made *no gain* from the first lesson to the second, and those who had received high scores showed high interest while those who received low scores did not. The results are summarized in table 3.

	Achievement gain	Attitude toward subject
Scores only	none	High achievers: positive Low achievers: negative
Comments only	30%	All positive
Scores and comments	none	High achievers: positive Low achievers: negative

Table 3: Impact of feedback interventions on achievement and attitude in Butler (1988)

Far from producing the best effects of both kinds of feedback, giving scores alongside the comments completely washed out the beneficial effects of the comments. When given both scores and comments, most students looked at the score first. The next thing most of these students looked at was the score of their nearest neighbor. In other words, if teachers write careful diagnostic comments on a student's work, and then put a score or grade on it, they are likely to be wasting the time they spend writing comments. The students who get the high scores do not need to read the comments, and the students who get the low scores do not want to. And yet the use of both scores and comments is probably the most widespread form of feedback used in the USA, and yet this study (and others like it—see below) show that it is no more effective than scores alone. The teacher would be better off just giving a score. The students will not learn anything from this but the teacher will save herself a great deal of time.

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

In two of the classes, students were given written comments;
In two of the classes, students were given grades;
In two of the classes, students were given written praise; and
In two of the classes, students were given no feedback at all.

The quality of the work done in the second lesson was compared to that done in the first. The quality of work of those given comments had improved substantially compared to the first lesson, but those given grades and praise had made no more progress than those given no feedback. At the end of the second lesson, the students completed a questionnaire about the factors that influenced their work which attempted to elicit from the students the reasons for their level of motivation. In particular the questionnaire sought to establish whether the students' effort investment was due to ego-related concerns (such as class rank) or task-related concerns (such as learning the material). The students who had received comments during their work on the topic had high levels of task-involvement, but their levels of ego-involvement were the same as those given no feedback. However, while those given grades and those given written praise had comparable levels of task-involvement to the control group, their levels of ego-involvement were substantially higher. The only effect of the grades and the praise, therefore, was to increase the sense of ego-involvement without increasing achievement. These findings are consistent with those of Cameron & Pierce (1994), who found that while verbal praise and supportive feedback did increase

students' interest in and attitude towards a task, such feedback had little if any effect on performance.

These findings are consistent with the research on praise carried out in the 1970s which showed clearly that praise was not necessarily "a good thing"—in fact the best teachers appear to praise slightly less than average (Good and Grouws, 1975). It is the quality, rather than the quantity of praise that is important. In particular, teacher feedback is far more effective when it is infrequent, credible, contingent, specific, and genuine (Brophy, 1981), and focuses on features of the work that are within the students' control (Siero and Van Oudenhoven, 1995; Dweck, 2000).

It is clear that from a young age, students' attitudes to learning are shaped by the feedback they receive. In a year-long study of eight kindergarten and first grade classrooms in six schools, Tunstall & Gipps (1996a; 1996b) identified a range of roles that feedback played in classrooms in England. Like Torrance & Pryor (1998), they found that a substantial proportion of the feedback given to students focused on socialization: "I'm only helping people who are sitting down with their hands up" (p. 395). Beyond this socialization role, they identified four types of feedback on academic work, which they labeled A, B, C and D as shown in table 4. In the first type was placed feedback that rewarded or punished the students for their work, for example when students were allowed to leave the classroom for lunch early if they had done good work, or where students were told that if they hadn't completed assigned tasks, they would have to stay in during lunch to finish their work. Type B feedback was, like type A, evaluative, but rather than focusing on rewards and sanctions, the feedback indicated the teacher's approval or disapproval (e.g. "I'm very pleased with you" vs. "I'm very disappointed in you today").

In addition, to the kinds of *evaluative* feedback described above, they identified two types of *descriptive* feedback. Type C feedback focused on academic work as product, and emphasized the adequacy of the work with respect to the teacher's criteria for success. At one end of the continuum, feedback focused on the extent to which the work already satisfied the criteria, while at the other end, the feedback focused on what the student needed to do to improve the work. An example of the former is, "This is extremely well explained" pointing out the way that the student has satisfied the teacher's criteria. An example of the latter is, "I want you to go over all of them and write your equals sign in each one"—here the emphasis is on what needs to be done to improve the work.

In contrast to the idea of work as product, Type D feedback focused on the process aspects of work, with the teacher cast in the role of facilitator, rather than judge or coach. As Tunstall and Gipps (1996a) explain, teachers engaged in this kind of feedback "conveyed a sense of work in progress, heightening awareness of what was being undertaken and reflecting on it" (p. 399).

Evaluative feedback	Type A	Type B
Positive	Rewarding	Approving
Negative	Punishing	Disapproving

Descriptive feedback	Type C	Type D
Achievement feedback	Specifying attainment	Constructing achievement
Improvement feedback	Specifying improvement	Constructing the way forward

Table 4: Typology of teacher feedback from Tunstall & Gipps (1996a)

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

The effects of feedback highlighted above might suggest that the more feedback, the better, but this is not necessarily the case. Day and Cordón (1993) looked at the learning of a group of 64 third grade students on reasoning tasks. Half of the students were given a "scaffolded" response when they got stuck—in other words they were given only as much help as they needed to make progress, while the other half were given a complete solution as soon as they got stuck, and then given a new problem to work on. Those given the "scaffolded" response learned more, and retained their learning longer than those given full solutions. In a sense, this is hardly surprising, since those given the complete solutions had the opportunity for learning taken away from them. As well as saving time for the teacher, developing skills of "minimal intervention" promote better learning.

A good example of such feedback is given by Saphier (2005 p. 92):

> Teacher: "What part don't you understand?"
>
> Student: "I just don't get it."
>
> Teacher: "Well, the first part is just like the last problem you did. Then we add one more variable. See if you can find out what it is, and I'll come back in a few minutes."

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

laid out, and this "busy work" can provide time for the student to make sense of the task herself (see discussion of performance versus mastery goals below).

The consistency of these messages from research on the effects of feedback extends well beyond school and other educational settings. A review by Kluger & DeNisi (1996) of 131 well-designed studies in educational and workplace settings found that, on average, feedback did improve performance, but this average effect disguised substantial differences between studies. Perhaps most surprisingly, in 40% of the studies, giving feedback had a negative impact on performance. In other words, in two out of every five carefully-controlled scientific studies, giving people feedback on their performance made their performance worse than if they were given no feedback on their performance at all! On further investigation, the researchers found that feedback made performance worse when it was focused on the self-esteem or self-image (as is the case with grades and praise). The use of praise can increase motivation, but then it becomes necessary to use praise all the time to maintain the motivation. In this situation, it is very difficult to maintain praise as genuine and sincere. In contrast, the use of feedback improves performance when it is focused on what needs to be done to improve, and particularly when it gives specific details about *how* to improve (see section on *Activating students as owners of their own learning* below).

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

Clarifying and sharing learning intentions and success criteria with learners

In a chapter entitled "The view from the student's desk" Mary Alice White (1971) suggested that students often had no real idea where they were going in their learning.

> The analogy that might make the student's view more comprehensible to adults is to imagine oneself on a ship sailing across an unknown sea, to an unknown destination. An adult would be desperate to know where he [sic] is going. But a child only knows he is going to school...The chart is neither available nor understandable to him... Very quickly, the daily life on board ship becomes all important ... The daily chores, the demands, the inspections, become the reality, not the voyage, nor the destination. (White, 1971, p. 340)

Certainly many authors, notably Bernstein (1975), have shown that not all students share the teacher's understanding of what they are meant to be doing in classrooms. As a simple example, consider the task of identifying the "odd one out" in the following list of objects: knife, fork, hammer, bottle of tomato ketchup. Some students believe that the bottle of tomato ketchup is the odd one out, since the others are all metal tools, while other students believe that the hammer is the odd one out since the others appear on the table at meal-times. Neither answer is wrong or right, of course, but as Keddie (1971) pointed out over 35 years ago, schools value some kinds of knowledge more than others, although what, precisely is desired or valued is not always made clear.

For example, in a study of 72 students between the ages of 7 and 13, Gray and Tall (1994) found that the reasoning of the higher-achieving students was qualitatively different from that of the lower-achieving students. This, of course, is not surprising, as the research on expertise (Berliner, 1994) shows that this is a characteristic of expertise. What was surprising in Gary and Tall's study was in the nature of the difference. In particular, the higher-achieving students were able to work with unresolved ambiguities about whether mathematical entities are concepts or procedures. In contrast, lower-attaining students, by refusing to accept the ambiguities inherent in mathematics are, in fact, attempting a far more difficult form of mathematics, with a far greater cognitive demand.

A simple example may be illustrative here. Consider the number $6\frac{1}{2}$. The mathematical operation between the 6 and the $\frac{1}{2}$ is actually addition. $6\frac{1}{2}$ is a shorthand expression for $6+\frac{1}{2}$ but when we write an expression in algebra by concatenating two terms, such as $6x$ the implied operation between the 6 and the x is multiplication, not addition. The relationship between the 6 and the 1 in 61 is different again. And yet, very few people who are successful in mathematics are aware of these inconsistencies or differences in mathematical notation. In a very real sense, being successful in mathematics requires knowing what to worry about and what not to worry about. Students who do not understand what is important and what is not important will be at a very real disadvantage.

Quite how big a difference this can make is brought out in a study of seventh grade science classes conducted by White and Frederiksen (1998). The study involved three teachers, each of whom taught 4 parallel seventh grade classes in two U.S. schools. The average size of the classes was thirty-one. In order to assess the representativeness of the sample, all the students in the study were given a basic skills test, and their scores were close to the national average. All twelve classes followed a novel curriculum (called ThinkerTools) for 14 weeks. The curriculum had been designed to promote thinking in the science classroom through a focus on a series of seven scientific investigations (approximately two weeks each). Each investigation incorporated a series of evaluation activities. In two of each teacher's four classes, these evaluation episodes took the form of a discussion about what they liked and disliked about the topic. For the other two classes, they engaged in a process of "reflective assessment." Through a series of small-group and individual activities, the students were introduced to the nine assessment criteria (each of which was assessed on a 5-point scale) that the teacher would use in evaluating their work. At the end of each episode within an investigation, the students were asked to assess their performance against two of the criteria. At the end of the investigation, students had to assess their performance against all nine. Whenever they assessed themselves, they had to write a brief statement showing which aspects of their work formed the basis for their rating. At the end of each investigation, students presented their work to the class, and the students used the criteria to give each other feedback.

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

The mean project scores achieved by the students in the two groups over the seven investigations are summarized in table 5, classified according to their score on the basic skills test.

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

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

Table 5: Mean project scores for students

Two features are immediately apparent in these data. The first is that the mean scores are higher for the students doing "reflective assessment," when compared with the control group—in other words, all students improved their scores when they thought about what it was that counted as good work. However, much more significantly, the difference between the "likes and dislikes" group and the "assessment" group was much greater for students with weak basic skills. This suggests that, at least in part, low achievement in mathematics is exacerbated by students' not understanding what it is they are meant to be doing.

Now although it is clear that students need to understand the standards against which their work will be assessed, the study by White and Frederiksen shows that the criteria themselves are only the starting point. At the beginning, the words do not have the meaning for the student that they have for the teacher. Just giving "quality criteria" or "success criteria" to students will not work, unless students have a chance to see what this might mean in the context of their own work. Furthermore, as Stiggins (2001 pp. 314-322) points out, there is a very real tension between task-specific and generic rubrics. Task-specific rubrics can be written in very clear language, and thus can communicate accurately to students what is required, but students need to come to grips with a new rubric for each task. Generic rubrics, on the other hand, can span a number of tasks, but are unlikely to provide clear guidance to a student about how to improve her or his work. Another weakness of task-specific rubrics is that, by definition, they focus only on the qualities needed for the specific task, and therefore may not contribute to more generalizable skills. Too much attention to one's performance relative to a detailed rubric may, in fact, be counterproductive (Kohn, 2006). For this reason, Arter & McTighe (2001) suggest that task-specific rubrics are more appropriate for summative assessment, where consistency of scoring is paramount, but that generic rubrics are more appropriate for formative assessment, since they focus on qualities that transcend the immediate task (for more on the characteristics of good rubrics, see Relearning by Design, 2000).

All the features of good rubrics discussed above apply to the use of rubrics for both summative and formative purposes, but if rubrics are to function formatively, additional considerations apply. Effective summative assessment requires those engaged in assessment to share the same construct of quality, and rubrics can play an important role in promulgating common standards (Sadler, 1987). Effective formative assessment requires the student to share the same construct:

> The indispensable conditions for improvement are that the student comes to hold a concept of quality roughly similar to that held by the teacher, is continuously able to monitor the quality of what is being produced during the act of production itself, and has a repertoire of alternative moves or strategies from which to draw at any given point. (Sadler 1989, p. 121)

However, it is possible for a rubric to identify accurately different degrees of quality in a piece of work without explicitly identifying the relationship between the levels. Both a teacher and a student may be able to identify that a particular piece of work merits a "4" on a six-point rubric, but be unclear about what needs to be done to move the piece of work forward to a "5" or a "6," much in the same way as we can tell a softball pitcher that she needs to lower her ERA or get her rising fast-ball to rise without having any idea how to accomplish this. For this reason, it is particularly helpful for formative purposes if rubrics are longitudinal or developmental, so that they are not just descriptions of quality, but effectively, *anatomies* of quality.

In clarifying and sharing learning intentions and success criteria with learners, it is also important to be clear about whether the particular aspect of quality sought can be communicated with a rubric at all. Some aspects of quality, such as whether calculations are correct, or whether diagrams are drawn in pencil and labeled, can easily be set down in a rubric. On the other hand, some features, such as whether an approach to a mathematical investigation is systematic or not cannot be so easily described. Using a term like "Adopts a systematic approach" in a rubric may look like a definition of quality, but it is unlikely to be meaningful to those who do not already know what it means to be systematic. In Michael Polanyi's terms, it is not a rule that can be easily put into effect, but is rather a *maxim*:

> Maxims cannot be understood, still less applied by anyone not already possessing a good practical knowledge of the art. They derive their interest from our appreciation of the art and cannot themselves either replace or establish that appreciation (Polanyi, 1958 p. 50).

In such situations, the best we can do is to help students develop what Guy Claxton calls a "nose for quality" (Claxton, 1995). Rubrics may have a role to play in this process, as they did in the study by White and Frederiksen. Rubrics were shared with students, but the students were given time to think through, in discussion with others, what this might mean in practice, applied to their own work. We shouldn't assume that the students will understand these right away, but the criteria will provide a focus for negotiating with students about what counts as quality in the mathematics classroom.

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

Activating students as owners of their own learning

The power of getting students to take some ownership of their own work is shown very clearly in an experiment by Fontana and Fernandez (1994). A group of 25 Portuguese primary school teachers met for two hours each week over a twenty-week period during

which they were trained in the use of a structured approach to student self-assessment. The approach to self-assessment involved an exploratory component and a prescriptive component. In the exploratory component, each day, at a set time, students organized and carried out individual plans of work, choosing tasks from a range offered to them by the teacher, and had to evaluate their performance against their plans once each week. The progression within the exploratory component had two strands—over the twenty weeks, the tasks and areas in which the students worked were to take on the student's own ideas more and more, and secondly, the criteria that the students used to assess themselves were to become more objective and precise.

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

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

Another 20 teachers, matched in terms of age, qualifications, experience, using the same curriculum scheme, for the same amount of time, and receiving the same amount of professional development, acted as a control group. The 354 students being taught by the 25 teachers using self-assessment, and the 313 students being taught by the 20 teachers acting as a control group were each given the same mathematics test at the beginning of the project, and again at the end of the project 20 weeks later. Over the course of the experiment, the scores of the students taught by the control-group teachers improved by 7.8 points. The scores of the students taught by the teachers developing self-assessment improved by 15 points—almost twice as big an improvement. In other words, the students participating in the self-assessment learned in 20 weeks what the students in the control group classrooms would take 40 weeks to learn—a doubling of the rate of learning.

Whether the students in this study were able to assess their own performance objectively is not clear, and in general, such questions are a matter of heated debate, but very often the debate takes place at cross-purposes. Opponents of self-assessment say that students cannot possibly assess their own performance objectively, but this is an argument about the *summative* function of self-assessment, which is beyond the scope of this chapter. The focus here is whether activating students as owners of their own learning does, indeed, enhance learning, and in this regard, concordance with other, external, judgments is a secondary concern. What matters is whether enhancing students' engagement with, and ownership of their own learning enhances learning. For this to happen students have to be motivated (literally, moved) to do so, but they also have to possess the necessary cognitive resources.

The key concept here is *self-regulation*, defined by Boekaerts, Maes and Karoly (2005) as "a multilevel, multicomponent process that targets affect, cognitions, and actions, as well as

features of the environment for modulation in the service of one's goals" (Boekaerts, 2006 p. 347). Put simply, students who lack self-regulation skills are unlikely to be able to take control of and guide their own learning without the capability for self-regulation (DeCorte, Verschaffel & Op't Eynde, 2000). Inevitably, with such a broad area, different researchers have emphasized different aspects of self-regulation.

Winne (1996) emphasized the cognitive aspects of this process, defining self-regulated learning as a ""metacognitively governed behavior wherein learners adaptively regulate their use of cognitive tactics and strategies in tasks" (p. 327). Others have observed that many students appear to possess the necessary skills of self-regulation, but do not use them in classrooms (and especially in mathematics classrooms), which suggests that the problem in not a lack of skill, but rather a lack of motivation or volition (Corno, 2001). Still others have argued for the integration of different perspectives within sociocultural (McCaslin & Hickey, 2001; Hickey & McCaslin, 2001) or social constructivist (Op't Eynde, DeCorte & Verschaffel, 2001) perspectives.

These two broad areas—cognition and motivation—have been extremely fertile areas of inquiry over the last quarter century or so, but unfortunately, most of the research undertaken has been firmly in one area or the other (Wigfield, Eccles & Rodriguez, 1998). As Sorrentino & Higgins (1986) point out there has been an "almost total exclusion of motivation in research on cognition" (p. 3) and Nisbett and Ross (1980), in their influential book on human reasoning, lamented psychology's "inability to bridge the gap between cognition and behavior" (p. 11). If we are to realize the potential for self-regulation to improve learning, however, ways of bridging this gap must be found, because, as Boekaerts (2006) shows, self-regulated learning is both metacognitively governed *and* affectively charged (p. 348).

In the next two sections, the research on cognition and motivation will be briefly reviewed, focusing on the aspects that are most relevant to the idea of activating students as owners of their own learning. The implications of this research for using assessment to activate students as owners of their own learning will then be explored through the use of the dual-processing model of self-regulation proposed by Boekaerts (2006).

Metacognition

It is perhaps not too much of an overstatement to say that everyone agrees that metacognition is important, but no-one can agree on what it is, at least in a sufficiently precise definition to put into practice. Even John Flavell, widely credited with inventing the term, acknowledged that it was a "fuzzy concept" (Flavell, 1981 p. 37). Literally, it is "beyond thinking" or thinking about thinking. Flavell (1976) defined the concept thus:

> "Metacognition" refers to one's knowledge concerning one's own cognitive processes and products or anything related to them, e.g., the learning-relevant properties of information and data. For example I am engaging in metacognition (metamemory, metalearning, metaattention, metalanguage, or whatever) if I notice that I am having more trouble learning A than B; if it strikes me that I should double-check C before accepting it as a fact; if it occurs to me that I had better scrutinise each and every alternative in any multiple-choice type task situation before deciding which is the best one; if I sense that I had better make a note of D because I may forget it; if I think to ask

someone about E to see if I have it right. In any kind of cognitive transaction with the human or nonhuman environment, a variety of information processing activities may go on. Metacognition refers, among other things, to the active monitoring and consequent regulation and orchestration of these processes in relation to the cognitive objects or data on which they bear, usually in the service of some concrete goal or objective. (p. 232)

It is thus an "umbrella term" that encompasses all "knowledge and cognition about cognitive phenomena" (Flavell, 1979 p. 906), including knowing what one knows ("metacognitive knowledge"), what one can do ("metacognitive skills") and what one knows about one' own cognitive abilities ("metacognitive experience"). Kluwe (1982 p.202) broadened the term further, by including knowledge about the thinking of others as well as oneself. Brown (1987) identifies "four historically separate, but obviously interlinked, problems in psychology that pertain to issues of metacognition" (p. 69):

- whether individuals can have access to their own cognitive processes
- who or what is responsible for executive control of cognition
- what is the role of self-regulation within learning
- what is the role of regulation by others in learning

although there is a continuing debate about whether metacognition has to be a conscious process (Carr, Alexander & Folds-Bennett, 1994).

Studies have repeatedly shown that students with greater awareness of their own cognitive processes have higher achievement (DeCorte, Verschaffel & Op't Eynde, 2000), as do students who display more conscious self regulation (see for example, Stillman & Galbraith, 1998). Butler & Winne (1995) summarized the situation thus: "Theoreticians seem unanimous—the most effective learners are self-regulating" (p. 245), but this leaves unresolved the question of whether improving metacognitive skills improves achievement, and this is where the lack of an agreed operational definition of metacognition is most serious.

Many studies have shown that training students in metacognitive strategies improves their performance even in the early schools years (e.g. Lodico, Ghatala, Levin, Pressley & Bell, 1983) and can also generalize metacognitive strategies to new situations (see Hacker, Dunlosky & Graesser, 1998 for a review), but other studies have failed to find benefits for metacognitive training. For example, in a multi-level study of 444 seventh grade students in 18 heterogeneous classrooms in the Netherlands, Hoek, van den Eeden & Terwel (1999) found that the achievement of students trained in metacognitive and social strategies simultaneously was no greater than controls just told to work together (although differences in two of the three outcome measures were found in single-level models). Without a clear operational definition of metacognition, this result is difficult to interpret. Those who advocate the benefits of metacognitive training can just claim that the unsuccessful efforts were not "real" metacognition.

What is clear is that many studies *have* found clear benefits for metacognitive interventions in real classroom settings as well as laboratories. In the previous edition of this handbook, Schoenfeld (1992) reviewed the research on metacognition with a specific focus on the learning of mathematics, although most of the studies available at that time had been conducted in laboratory settings, or in small-scale trials. Since then, a series of studies have

shown that training in self-evaluation improves student achievement in real settings, and across extended periods of time.

For example, Ross, Hogaboam-Gray & Rolheiser (2002) reported on a study in which 12 fifth and sixth-grade mathematics teachers were trained in the promotion of systematic student self-evaluation, and the performance of their students (N = 259) was compared with students (N = 257) taught by teachers who had not been so trained. After a twelve-week period, the students undertaking self-evaluation outscored the other students by 0.4 standard deviations on a mathematics task.

Mevarech and her colleagues in Israel have demonstrated that metacognitive training improves student achievement in a range of settings, and across even longer periods of time. Their method is termed "IMPROVE" after the initial letters of the key components:

Introducing the new concepts;
Metacognitive questioning in small groups;
Practicing, **R**eviewing, and reducing difficulties
Obtaining mastery;
Verification and **E**nrichment

In one year-long study 99 seventh-grade students in three classrooms in Israel were taught using the IMPROVE method, and their performance was compared with 148 students from five other classes (Mevarech & Kramarski, 1997). Although the IMPROVE students scored slightly lower than the control group students on a pre-test, they outscored controls on both an "Introduction to algebra" test and on a test of mathematical reasoning. The standardized effect size *d* (Cohen, 1988) of the advantage for the IMPROVE students was 0.31 standard deviations for the "Introduction to algebra" test and 0.44 standard deviations on the test of mathematical reasoning. A second study involving 265 seventh-graders found a similar result (d=0.38).

Other studies found that the IMPROVE method was more effective than just training in metacognitive skills (Kramarski, Mevarech & Lieberman, 2001) or structured collaborative work (Kramarski, Mevarech & Arami, 2002) and the benefits extended to both traditional measures of achievement and authentic tasks. More recent work has established that achievement gains are still present after one year, and are larger for lower-achieving students (Mevarech & Kramarski, 2003).

Another research program—Cognitive Acceleration in Mathematics Education or CAME (Adhami, Johnson & Shayer, 1998)—also aims to increase student achievement in mathematics by attention to metacognitive processes. Although it shares many features with IMPROVE, the program appears to make somewhat greater demands on teachers and is therefore more difficult to implement than the IMPROVE project. Results from 68 classrooms monitored over a two year period suggest that the average increase in student achievement was approximately 0.34 standard deviations (Adhami, Johnson & Shayer, 1997), although the distribution was trimodal, with one peak around 0.8 standard deviations (almost all the classes coming from one school), one peak around 0.5 standard deviations, and another at 0 (which appears to be attributable to schools that did not implement the program as intended).

Other cognitive-behavior interventions have been reviewed by Boekarts & Corno (2005 pp. 213-221), but the lack of an adequate theorization of metacognition makes comparing different interventions difficult, if not impossible. Nevertheless, the results from IMPROVE, CAME, and similar interventions suggest that metacognitive interventions can have substantial impact on student achievement in mathematics, even if the design of such interventions is currently more art than science. It is also no coincidence that the most effective programs for developing metacognitive skills are firmly rooted in specific domains of knowledge (Bransford, Brown & Cocking, 2000). However, just possessing the skills of being able to monitor one's own learning is not enough for students to learn mathematics. They also have to want to do so, and, as any mathematics teacher is all too aware, this cannot be taken for granted.

Motivation and learning

As Edward Deci (1996) notes, "Most people seem to think that the most effective motivation comes from outside the person, that it is something that one skilful person does to another" (p. 9), although there is substantial evidence that the situation is not that simple. In their explanation of Self-Determination Theory (SDT), Deci & Ryan (1985), distinguished between different types of motivation based on the different reasons or goals that give rise to a particular action. "The most basic distinction is between *intrinsic motivation*, which refers to doing something because it is inherently interesting or enjoyable, and *extrinsic motivation*, which refers to doing something because it leads to a separable outcome" (Ryan & Deci, 2000 p. 55, emphases in original).

Because of its power to drive quality learning, the notion of intrinsic motivation has been extensively studied, especially how the actions of teachers and parents can build, or undermine, it (Ryan & Stiller, 1991; Connell and Wellborn, 1991), and the basic distinction between intrinsic and extrinsic motivation has held up pretty well. However, while the benefits of intrinsic motivation are clear, it is far from clear that extrinsic motivation is antithetical to learning. For example, a meta-analysis of 61 studies that compared a group receiving rewards with a group that did not found that the effects were, on average, very small. Unanticipated verbal rewards increased both attitudes towards tasks, and the time that participants chose to spend on them. Anticipated rewards had no impact on attitude, but did reduce time spent on tasks, with a much stronger effect when the reward was independent of the quality of performance (Eisenberger and Cameron, 1996). As Ryan & Deci (2000) themselves point out, while intrinsic motivation is by definition autonomous, the fact that motivation is extrinsic does not mean that the motivation is necessarily *not* autonomous:

> For example, a student who does his homework only because he fears parental sanctions for not doing it is extrinsically motivated because he is doing the work in order to attain the separable outcome of avoiding sanctions. Similarly, a student who does the work because she personally believes it is valuable for her chosen career is also extrinsically motivated because she too is doing it for its instrumental value rather than just because she finds it interesting. (Ryan and Deci, 2000 p. 60)

In these two cases, neither student is intrinsically motivated, but the first is primarily just complying with an external control, while the second has embraced the task and is working because she values the goal. In other words, what controls the behavior of the individual involves both the motivation itself is internal or external, and whether the value system is internal or external.

When both are external, then behavior is *externally* regulated "by contingencies overtly external to the individual" Deci & Ryan, 1994 p.6), while *introjected* regulation "refers to behaviours that are motivated by internal prods and pressures such as self-esteem-relevant contingencies" (p.6), such as the student described above doing his homework under threat of sanctions. *Identified* regulation "results when a behaviour or regulation is adopted by the self as personally important or valuable" (p.6), as in the case of the student doing the work for her career, while *integrated* regulation "results from the integration of identified values and regulations into one's coherent sense of self" (p.6). These four types of extrinsic motivation—external, introjected, identified and integrated—form a continuum, in that their inter-correlations have been shown to conform to a simplex pattern in which the largest intercorrelations are closest to the leading diagonal (Ryan & Connell, 1989) and so, together with intrinsic motivation and amotivation (where the individual lacks any intention to act), provide a taxonomy of six kinds of motivation. While there is undoubtedly a clear hierarchy in terms of the degree of self-determination, this is clearly not a developmental continuum: a student may start out intrinsically motivated, but maintain activity only because of external sanctions or instrumental goals, while someone may engage in an activity because of the rewards associated with it, but then maintain the activity even when rewards are withdrawn (Eisenberger & Cameron, 1996).

The research on internal and external drivers of motivation accounts well for a range of observed student behaviors, but pays relatively little attention to why an individual learner might find something interesting or not. Hidi & Harackiewicz (2000) define interest as "an interactive relation between an individual and certain aspects of his or her environment (e.g., objects, events ideas), and is therefore content specific" (p. 152) involving both cognitive and affective components. They point out that researchers have placed different emphases on different aspects of interest, with some focusing on specific stimuli that focus attention which may or may not last (generally termed *situational* interest) while others have focused on more stable motivational orientations or personal dispositions towards particular topics or domains (termed *individual* or *personal* interest). As Dewey (1913), noted, catching and holding student interest are two very different things. Mitchell (1993) found that it was possible to generate student interest in mathematics through the use of group work, puzzles, and computer-based activities, but that such interest waned over time. In contrast, meaningfulness and involvement in the tasks built interest that *was* sustained over time.

The approach to motivation discussed above seeks to examine motivation as a cause of engagement in activity. This has been very powerful in understanding why people, and in particular, students in mathematics classes, do what they do. However, Csikszentmihalyi (1990) turns this on his head by looking at what we describe as "motivation" as an *emergent* property of the interaction between the task in which the individual is engaged, and the competences that the individual brings to the task. In this view, motivation is therefore not the cause, but the consequence of engagement. When the task demand is high, and the skill levels is low, then the individual experiences anxiety, while if the task demand is low and the skill level is high, the individual experiences boredom. But if the task demand is just at the limit of the skill level, then the individual experiences "flow:"

> A dancer describes how it fees when a performance is going well: "Your concentration is very complete. Your mind isn't wandering, you are not thinking of something else; you are totally involved in what you are doing.…Your energy is flowing very smoothly. You feel relaxed, comfortable and energetic."
>
> A rock climber describes how it feels when he is scaling a mountain: You are so

> involved in what you are doing [that] you aren't thinking of yourself as separate from the immediate activity….You don't see yourself as separate from what you are doing."
>
> A mother who enjoys the time spent with her small daughter: "Her reading is the one thing she's really into, and we read together. She reads to me and I read to her, and that's a time when I sort of lose touch with the rest of the world, I'm totally absorbed in what I'm doing."
>
> A chess player tells of playing in a tournament: "…the concentration is like breathing—you never think of it. The roof could fall in and, if it missed you, you would be unaware of it." (Csikszentmihalyi, 1990 pp. 53-54)

As well as the drivers of interest that Hidi & Harackiewicz identify, therefore, it is also necessary to attend to matching the level of demand to the skill of the individual.

Of course, the research reviewed above represents somewhat of a "counsel of perfection." Ideally we would want all students to be intrinsically motivated, and to experience "flow" in their mathematics classrooms, but this is impossible to achieve in even the best mathematics classrooms all the time. We therefore need to take account of why students do what they do even if they are not intrinsically motivated.

Eccles, Adler, Futterman, Goff, Kaczala, Meece & Midgley, (1983) identify four components of task value: attainment value, intrinsic value, utility value and cost. Attainment value refers to the value to an individual of doing well on a task, particularly in terms of confirming or disconfirming particular aspects of one's identity (see also Boaler, Wiliam & Zevenbergen, 2000). Intrinsic value corresponds to the ideas of intrinsic motivation and flow discussed above. Utility value encompasses the value of the task in terms of current and future goals, and is closely related to the identified, introjected and external forms of regulation described by Deci and Ryan above. Finally, Eccles *et al*. (1983) emphasize the importance of a trade-off between the perceived value (whether it is attainment value, intrinsic value or utility value) of a task and its cost. Cost includes both the opportunity cost that attempting a task might take, and also the negative consequences such as performance anxiety, risk to one's view of self if unsuccessful, and so on.

The goals that students actually pursue in mathematics classrooms will therefore depend on a complex calculus of cost and benefit. Bandura (1986) and Schunk (1990; 1991) have shown that students are more motivated to reach goals that are specific, within reach, and offer some degree of challenge, but more recently, researchers have focused on broader issues, and in particular whether the student's primary orientation is towards *mastery* or *performance*. Ames (1992) defined an achievement goal as follows:

> An *achievement goal* concerns the purposes of achievement behavior. It defines an integrated pattern of beliefs, attributions, and affect that produces the intentions of behavior and that is represented by different ways of approaching, engaging in, and responding to achievement activities. (p. 261, emphasis in original)

Mastery achievement goals are those related to the acquisition of new skills or increasing the level of competence, while performance achievement goals are those that are related to the level of performance one is able to produce. Students with mastery goals seek to increase their level of competence, while those with performance goals are motivated either to show what they can do (performance approach goals) or to avoid failing (performance avoidance goals). While mastery

goals are clearly beneficial, and performance-avoidance goals are clearly detrimental to learning, there is still much debate about the usefulness of performance-approach goals. In a study of a 5-week elementary school math unit, Linnenbrink (2005) found that a combined approach, emphasizing mastery goals within small groups but emphasizing performance goals between groups produced the highest achievement, provided the competition between groups was focused on relative improvement amongst the groups (see also the discussion on *Activating students as instructional resources for one another* below).

Whether students choose to pursue mastery or performance goals depends, of course, on a range of factors, including the student's beliefs about the likelihood of success, and this, in turn, is influenced by students' ideas about the causes of success in school. Through an extensive research program going back over a quarter of a century, involving questionnaires and interviews with thousands of students, Dweck (2000) has shown that there are substantial differences between students in their beliefs about the causes of success and failure in the classroom. Two dimensions appear to be particularly important:

Personalization: whether success is due to "internal" factors (such as one's own performance) or "external" factors (such as getting a lenient or severe marker);

Stability: being due to "stable" factors (such as one's ability) or "unstable" factors (such as effort or luck);

Dweck and others have found that many if not most students attribute success and failure differently. So for example, the student who says "I got an A" but "She gave me a C" is attributing success internally, but attributing failure externally. In fact many studies have found that boys are more likely to attribute their successes to internal, stable causes (such as ability), and their failures to external, unstable causes (such as bad luck or hostile teachers). This would certainly explain the high degree of confidence with which many boys approach tests or examinations for which they are completely unprepared (the important question for such students is not "Have I thoroughly reviewed the material for this test?" but "Do I feel lucky today?"). More controversially, the same research suggests that girls attribute their successes to internal unstable causes (such as effort) and their failures to internal stable causes (such as lack of ability), leading to what has been termed "learned helplessness." What is clear is that the most adaptive beliefs are that both success and failure are attributable to internal, unstable causes: "it's down to you, and you can do something about it." This is why the strategy of *Clarifying and sharing learning intentions and success criteria with learners* discussed above is so important. When students are clear about the criteria for success, they are more likely to attribute success and failure to internal unstable causes.

In the 1970s and 1980s, efforts to raise the confidence of students that they could achieve in mathematics focused on self-esteem as a critical factor, but Bandura (1977) has convincingly argued that self-efficacy—a generative capacity to carry one's plans through to completion—is a far more useful focus of inquiry. Students' expectations about their likelihood of success strongly influence their decisions about which tasks to attempt, how much effort to expend, and their persistence in the face of difficulties (Bandura, 1997), and increasing students' self-efficacy has been shown to increase performance in mathematics. Indeed, it appears that self-efficacy may be more important in mathematics than any other school subject (Pajares, 1996 p. 555).

Students' decisions about whether to pursue mastery or performance goals are also influenced by teachers' practices. In a study of 1571 students in 84 mathematics classrooms from fifth to twelfth-grade Deevers (2006) used hierarchical linear modeling to explore the relationship between teachers' formative assessment practices and student attitudes and beliefs about mathematics. He found that while students' self-efficacy beliefs and motivation to learn declined steadily from fifth to twelfth grade, students provided with positive constructive feedback were more likely to display mastery orientation, even though teachers gave less of this kind of feedback as students got older.

Closely related to the work on self-efficacy and goal orientation are studies undertaken by Carol Dweck and her colleagues into student perceptions and beliefs about the nature of ability in school (see Dweck, 2000 for a comprehensive summary of this work). Dweck & Leggatt (1986) showed that many students believe that ability is fixed: there are smart people and not so smart people. Student holding this "entity" view of ability are likely to adopt a performance orientation to challenging tasks that are set for them. If they are confident in their ability to achieve what is asked of them, then they will attempt the task. However, if their confidence in their ability to carry out their task is low, then, unless they believe that the task is so hard that no-one is expected to succeed, they will avoid the challenge. In short, these students are deciding that they would rather be thought lazy than dumb.

In contrast, there are other students who see ability as incremental. They see challenging tasks as chances to learn—to get smarter—and therefore in the face of failure will try harder (Mayer, Turner & Spencer, 1997). What is perhaps most important here is that these views of ability are generally not global—the same students often believe that ability in mathematics is fixed, while at the same time believe that ability in, for example, athletics is incremental, in that the more one trains, the more one's ability increases. The crucial thing, therefore, is that teachers inculcate in their students a belief that ability is incremental rather than fixed. Doing so would be advantageous even if it were not true, since the consequences would be so beneficial, but, as was discussed in the section on *Eliciting evidence* above, there is overwhelming evidence that "smart is not something you are; it's something you get."

Integrating motivational and cognitive perspectives

A number of ways of bringing together the motivational and cognitive perspectives on self-regulation have been proposed. One of the most promising is the "dual processing" theory developed by Boekaerts (1993). In the model:

> It is assumed that students who are invited to participate in a learning activity use three sources of information to form a mental representation of the task-in-context and to appraise it: (1) current perceptions of the task and the physical, social, and instructional context within which it is embedded; (2) activated domain-specific knowledge and (meta)cognitive strategies related to the task; and (3) motivational beliefs, including domain-specific capacity, interest and effort beliefs. (Boekaerts, 2006 p. 349)

As a result of the appraisal of the task, the student begins to act along one of two pathways. If the task appraisal is positive, the student "crosses the Rubicon" (Winne, 2005 pp. 234-236) and begins activity along the "growth pathway" where the goal is to increase

competence. In this case, the self-regulation is "top-down" in that the flow of energy is directed by the student. If, on the other hand, the task appraisal is negative, attention shifts away from the learning task and towards the well-being pathway. The student then becomes focused on self-appraisal rather than task appraisal, concentrating on preventing threat, harm or loss. This form of self-regulation is termed "bottom-up" by Boekaerts since it is triggered by cues in the environment, rather than by learning goals. Such "bottom-up" regulation is not necessarily negative, since by attending to the well-being pathway, the student may find a way to restore well-being, thus allowing a shift of energy and attention back to the growth pathway.

An important feature of the dual regulation model is that it hypothesizes that the balance between "top-down" and "bottom-up" pathways of regulation is dynamic, rather than being a dispositional feature of an individual student. Experimental support for this hypothesis was provided by a series of studies summarized in Boekaerts (2001 pp. 24-26) in which it was found that there was no direct link between domain-specific motivational beliefs and learning intention in any of the mathematics classrooms they studied. Rather, students' decisions about whether to invest effort in a mathematics assignment depended largely on their appraisal of the specific mathematics task in front of them. There is also evidence that interpretations of assessment outcomes from friends and parents mediate the relationship between the assessment outcomes themselves and how they impact students beliefs (Ross, Rolheiser & Hogaboam-Grey, 2002).

The dual processing model allows us to relate the various perspectives on cognition and motivation and learning described above. To simplify somewhat, students who are *personally*, as opposed to *situationally*, interested in a task are likely to engage in activity along the growth pathway, although for students who are not personally interested in a task, a range of factors related to the task-in-context may spark situational interest, thus also triggering activity along the growth pathway. Where personal interest is not the main driver of attention, more concern will be given to considerations of task *value* versus *cost*. The *integrated* and *identified* forms of regulation defined by Deci & Ryan are related to activity along the growth pathway, while *external* or *introjected* forms of regulation are related to activity along the well-being pathway. Students who display *mastery* orientation are activating the growth pathway and those displaying *performance* orientation are activating the well-being pathway.

Self-efficacy beliefs can drive progress along either pathway. Along the growth pathway, self-efficacy drives adaptive (meta)cognitive strategy use while along the well-being pathway, self-efficacy beliefs are likely to steer the learner away from *performance-avoidance goals* and towards *performance-approach goals*. Similarly views of ability as *incremental* help the learner stay on the growth pathway, while *entity* views of ability direct activity towards the well-being pathway, where details of the task-in-context, appraised in the light of views of personal capability, will influence decisions about whether to engage in the task or not.

This extended discussion of the research on metacognition, motivation and beliefs may appear to have taken us some way from the central theme of this chapter but an understanding of these issues is essential in developing an understanding of the role of assessment in learning. Adaptive adjustments to instruction are necessary to maximize learning, so assessment is an essential part of learning, but assessment processes themselves impact the learner's willingness, desire, and

capacity to learn (Harlen & Deakin-Crick, 2002). The research reviewed briefly above shows clearly that practices ostensibly designed to support learning actually prevent it from taking place (recall that the research reviewed by Kluger & DeNisi (1996) discussed above showed that in 40% of rigorously designed and conducted studies, feedback designed to improve performance actually had the opposite effect).

Although many issues are still unresolved, and a full theoretical synthesis of all the research perspectives is some way off if it is even possible, the existing research on cognition and motivation paints a reasonably coherent picture, and provides strong guidance on the design of classroom assessment environments (Brookhart, 1997) that can activate students as owners of their own learning. Feedback to learners should focus on what they need to do to improve, rather than on how well they have done, and should avoid comparison with others. Students who are used to having every piece of work scored or graded will resist this (see, for example, Smith & Gorard, 2005), wanting to know whether a particular piece of work is good or not, and in some cases, depending on the situation, the teacher may need to go along with this. In the long term, however, it seems that we should aim to reduce the amount of ego-involving feedback given to learners (and with new entrants to the school, perhaps not begin the process at all), and focus on the student's learning needs (Kohn, 1999). Furthermore, feedback should not just tell students to work harder or be "more systematic"—the feedback should contain a recipe for future action, for otherwise it is not formative. Finally, feedback should be designed so as to lead all students to believe that ability—even in mathematics—is incremental. In other words the more we "train" at mathematics, the smarter we get.

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

Activating students as instructional resources for one another

As Slavin, Hurley and Chamberlain (2003) note, "Research on cooperative learning is one of the greatest success stories in the history of educational research" (p. 177). However, despite the vast amount of research that has been done in this area, and the consensus that cooperative learning does increase student achievement, there is much less agreement on why cooperative learning is effective. Slavin et al. (2003) suggest that there are four major theoretical perspectives on cooperative learning and its effects on achievement.

Two of the perspectives—the motivational perspective and the social cohesion perspective—focus on the role of motivation. Adherents of the motivational perspective hold that students help their peers to learn because it is in their own interests to do so; the only way they can achieve their personal goals is to ensure that all the individuals in the group are successful. In contrast, theorists within the social cohesion perspective propose that students help their peers because they care about the group. In both of these perspectives, increased student achievement comes primarily from increased motivation and effort.

The other two perspectives—the developmental perspective and the cognitive elaboration perspective—emphasize the idea that increased learning in cooperative groups stems from the increased cognitive engagement produced in small groups. Empirical work in these two perspectives tends not to emphasize the importance of group goals, which are the hallmark of the motivational perspective, nor the building of team spirit, as required in the social cohesion perspectives. Indeed, many writers in this tradition, such as Damon (1984) have explicitly rejected the idea that extrinsic incentives are important in group learning situations (p. 337).

The developmental perspective draws strongly on the work of Piaget and Vygotsky, and in particular the latter, although as Shayer (2003) notes, the differences between these two are much less than is often assumed. The starting point for much of this work is Vygotsky's definition of the zone of proximal development (ZPD) as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978 p. 86)[iv]. Research in the developmental perspective emphasizes the role of more advanced peers in guiding the learning of others.

The fourth perspective—the cognitive elaboration perspective—like the developmental perspective, locates the benefit of cooperative learning in the interaction between students rather than in the motivation for engagement. However, while the developmental perspective predicts benefits for the recipient of the support, the cognitive elaboration perspective emphasizes that providers of support also benefit.

Of course, none of these perspectives is contradictory. Slavin et al. (2003) suggest a "logic model" in which group goals lead to both an increased personal motivation to learn as well as a motivation to ensure one's group-mates also learn. This in turn yields peer assessment and correction, with elaborated explanations, and peer modeling, resulting in enhanced learning. However, while the four perspectives may be complementary, there are marked differences in the degree of empirical support for each.

Slavin (1995) reviewed studies of cooperative learning in elementary and secondary schools and found 99 that involved interventions of at least four weeks' duration, and where the performance of students involved in cooperative learning was compared with students who were not. Of the 64 studies which provided *group* rewards that were contingent on the aggregate of the learning of *individual* members, over three quarters (50 out of 64) found significant positive effects on learning, and in none of the 64 studies was the effect of cooperative learning negative. The median effect size in these studies was 0.32 standard deviations, compared with a median effect size of 0.07 in the studies where there were no group rewards, or where the group rewards were based on a single group product.

In reviewing the available research on interventions designed to increase social cohesion, Slavin (1995) concluded that interventions that are designed to increase social cohesion, but do not provide rewards based on the learning of all group members, are no more effective than traditional instruction.

A review of 17 studies looking at the effects of peer interaction on the learning of mathematics, ranging from 2nd grade to 11th grade (Webb, 1991) found that those who provided help benefited little when this help was in the form of answers, procedural

information, or managerial information. However, when the feedback was in the form of elaborated explanations, those providing help benefited substantially. The median partial correlation between giving help in the form of elaborated explanations and achievement in mathematics, across the 10 studies on which the relevant information was available, was 0.27. In other words, a one standard deviation increase in giving elaborated help resulted in a 0.27 increase in achievement for the help-giver even after controlling for ability. The same study also found a strong negative effect of approximately the same magnitude for students who asked for help, but were given only the answer. The fact that these results were similar across topics and grades suggests that they are likely to generalize to a variety of mathematics classrooms.

Perhaps surprisingly, the effect of peer-tutoring can be almost as strong as one-to-one instruction from a teacher, and can be stronger than small-group instruction from a teacher, although some studies have found a strong correlation between the mathematical ability of the student giving help, the quality of explanation given, and the amount of learning that results (Fuchs, Fuchs, Karns, Hamlett, Dutka & Katzaroff, 1996). In a study by Shacter (2000) 109 students drawn from one fourth grade classroom, one fifth grade classroom and three combined fifth and sixth grade classrooms were randomly assigned to one of four treatments: teacher-student dyad, student-student dyad, same-age peer group with a teacher, and same-age peer group without a teacher. Compared with student-student pairs, the highest achievement was found in teacher-student pairs (standardized effect size of 0.53, 0.86 and 0.55 on concept mapping, essay and declarative knowledge respectively), but the achievement of students in student-led groups was almost as high (effect sizes 0.50, 0.57 and 0.30), and on average higher than the achievement in teacher-led groups (0.33, 0.67, 0.14). These effects may appear to be larger than some of the results reported above, but the measures in this study were much more sensitive to instruction than is typical for standardized tests.

In fact, a range of studies that have explored the effects of group rewards in combination with other interventions have consistently concluded that the "active ingredient" is the establishment of group goals and individual accountability. Fantuzzo, King and Heller (1992) compared the effects of reward and structure in the context of reciprocal peer teaching in mathematics. In one group, pairs of students were rewarded with a choice of activities if the sum of their scores on a test exceeded a threshold set by the teacher (reward only). In a second group students were trained in a structured method for tutoring each other, taking turns to be tutor and tutee (structure only). A third group of students received both interventions (reward+structure) while a fourth received neither (control). The results showed that reward had a greater effect on mathematics achievement than structure, but the combined effect was greater still. In another study, (Fantuzzo, Davis & Ginsburg, 1995) 72 fourth and fifth grade students in urban schools evidencing difficulties in mathematics were assigned randomly to one of three treatments. One-third of the students received a home-based parental involvement program, one third of the students received the parental involvement together with a reciprocal peer-tutoring intervention, and one third of the students were assigned to a control group where they just practiced work on which they were having difficulties. Students engaged in both reciprocal peer teaching and parental involvement were more confident and had higher mathematics achievement (both on classroom assessments and standardized tests) than either of the other two groups.

However, despite the clarity of the research evidence about the importance of group goals and individual accountability, it appears that while teachers acknowledge the importance of collaborative learning, few implement it in a way that the research suggests would be necessary for it to be effective. Antil, Jenkins, Wayne and Vadasy (1998) surveyed 85 elementary school teachers in two districts, and while 93% of the teachers said they employed collaborative learning, follow-up interviews with 21 of the teachers showed that only 5 teachers implemented collaborative learning in such a way as to create group goals together with individual accountability. Furthermore, only one of the 21 teachers implemented it collaborative learning that satisfied more complex criteria proposed by Cohen (1994): open-ended tasks that emphasize higher order thinking, group tasks that require input from other members, multiple tasks related to a central intellectual theme, and roles assigned to different group members. Whatever the research says about the benefits of peer collaboration, it would appear to be more difficult for teachers to implement than would at first appear.

Whether cooperative learning is more effective for some subgroups of students than others is still a matter of some debate. While some studies have found greater effects for higher attaining students (e.g. Stevens & Slavin, 1995), others have found more significant effects for lower-attaining students (e.g. Boaler, 2002). Slavin et al. (2003) conclude that most interventions appear to benefit high, medium and low achievers equally. There is, however, evidence that cooperative learning interventions are particularly beneficial for students of color. A meta-analysis of 37 studies on the effects of small group collaborative learning on achievement in post-secondary science, mathematics, engineering and technology (Springer, Stanne & Donovan, 1999) found a mean effect size of 0.51, with similar impact on persistence (d=0.46) and attitudes (d=0.55). The effects were larger for groups that were held outside class time (d=0.65) than for those held inside class time (d=0.44), and larger in four-year colleges (d=0.54) than two year colleges (d=0.21). The effect sizes on achievement were also larger for groups of African-American and Latina/o students (d=0.76) than for white (d=0.46) or heterogeneous groups (d=0.42). The particular value of collaborative learning environments for African-American students has also been emphasized by Boykin and his colleagues (Boykin, Coleman, Lilja & Tyler, 2004; Boykin, Lilja & Tyler, 2004).

The regulation of learning

The preceding sections have discussed in some detail the five key strategies of assessment for learning:

Clarifying and sharing 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
Activating students as instructional resources for one another
Activating students as owners of their own learning.

These five strategies can be integrated within a more general theoretical framework of the *regulation of learning processes* as suggested Perrenoud (1991, 1998)[v]. Within such a framework, the actions of the teacher, the learners, and the context of the classroom are all evaluated with respect to the extent to which they contribute to guiding the learning

towards the intended goal. In some environments, the responsibility for regulating learning, or "keeping learning on track" is with the teacher, in some it is with the student and in some it is shared (see Vermunt, 2003 for a classification of different kinds of learning environments in terms of the responsibility for regulation).

From this perspective, the task of the teacher is not necessarily to teach, but to create situations in which students learn. This focus emphasizes what it is that students learn, rather than what teachers do, although in an accountability-driven culture, it is hard to maintain this focus. One rather startling example of this is provided by a study of teachers being trained in teaching problem-solving skills to their students that was undertaken by Deci, Speigel, Ryan, Koestner & Kauffman (1982). Participating teachers were randomly allocated to one of two groups, and both groups were introduced to the important ideas about teaching problem-solving. In both groups teachers were given time to practise the problems, and were given both a list of useful hints and the actual solutions to all the problems. The only difference between the two groups was that at the end of the training session, one additional statement was made to the teachers in one group: "Remember, it is your responsibility as a teacher to make sure your students perform up to high standards." Subsequent analysis of recordings of the lessons of the teachers who had been told about the importance of "performing up to high standards" spent twice as much time talking during the teaching session as the other teachers and made three times as many directives and three times as many controlling statements (e.g., words like "should" and "must").

Obviously, regulating or controlling the activities in which students engage is only indirectly related to the learning that results (Clarke, 2001) and yet teachers appear to find it difficult to shift from planning activities to planning learning (Franke, Fennema & Carpenter, 1997). This is especially evident in interviews before lessons where teachers focus much more on the planned activities than on the resulting learning (e.g. "I'm going to have them do X"). In a way, this is inevitable, since only the activities can be manipulated directly. Nevertheless, it is clear that in teachers who have developed their formative assessment practices, there is a strong shift in emphasis away from regulating the activities in which students engage, and towards the learning that results (Black et al, 2003). Indeed, from such a perspective, even to describe the task of the teacher as teaching is misleading, since it is rather to "engineer" situations in which student learn.

However, in this context, it is important to note that the "engineering of learning environments" does not guarantee that the learning is proceeding in fruitful ways. Many visual arts classroom are *productive*, in that they do lead to significant learning on the part of students, but what any given student might learn is impossible to predict. An emphasis on the regulation of learning processes entails ensuring that the learning that is taking place is as intended.

When the learning environment is well-regulated, much of the regulation is pro-active, through the setting up of "didactical situations" (Brousseau, 1997). The regulation can be unmediated within such didactical situations, when, for example, a teacher "does not intervene in person, but puts in place a "metacognitive culture", mutual forms of teaching and the organization of regulation of learning processes run by technologies or incorporated into classroom organization and management" (Perrenoud, 1998 p100).

For example, a teacher's decision to use realistic contexts in the mathematics classroom can provide a source of proactive regulation, because then students can determine the

reasonableness of their answers. If students calculate that the average cost per slice of pizza (say) is $200, provided they are genuinely engaged in the activity, they will know that this solution is unreasonable, and so the use of realistic settings provides a "self-checking" mechanism. They are able to keep their learning on track themselves, rather than requiring a teacher's intervention, because they are "metacognitively, motivationally, and behaviorally active participants in their own learning process" (Zimmerman, 1986 p. 308)

On the other hand, the didactical situation may be set up so that the regulation is achieved through the mediation of the teacher or peers. McCaslin & Good (1996) proposed the term *co-regulation* to describe "the process by which the social instructional environment supports or scaffolds the individual via her relationships within the classroom, relationships with teachers and peers, objects and setting, and ultimately the self. Internalization of these supportive relationships empowers the individual to seek new challenges within co-regulated support" (p. 660). The teacher, in planning the lesson, can create questions, prompts or activities that evoke responses from the students that the teacher can use to determine the progress of the learning (such as the "hinge-point" questions described above), and if necessary, to make adjustments to the instruction. Examples of such questions are, "Is calculus exact or approximate?" or "Would your mass be the same on the moon?" (In this context it is worth noting that each of these questions is "closed" in that there is only one correct response—their value is that although they are closed, each question is focused on a specific conception.) In classroom discussion, through careful scaffolding of student discussion, teachers can develop the skills of self-regulation in their students (Mayer & Turner, 2002), especially through the establishment and maintenance of of sociomathematical norms (Yackel & Cobb, 1996) allowing students to support each other's learning in appropriate ways (Ross, 1995; McClain & Cobb, 2001).

Thoughtful "upstream" planning (before the lesson) therefore creates, "downstream" (during the lesson), the possibility that the learning activities may change course in the light of the students' responses. These "moments of contingency"—points in the instructional sequence when the instruction can proceed in different directions according to the responses of the student—are at the heart of the regulation of learning.

These moments arise continuously in whole-class teaching, where teachers constantly have to make sense of students' responses, interpreting them in terms of learning needs, and making appropriate responses. But they also arise when the teacher circulates around the classroom, looking at individual students' work, observing the extent to which the students are "on track." In most teaching of mathematics, the regulation of learning will be relatively tight, so that the teacher will attempt to "bring into line" all learners who are not heading towards the particular goal sought by the teacher—in these subjects, the goal of learning is generally both highly specific and common to all the students in a class. In contrast, when the class is doing an investigation, the regulation will be much looser. Rather than a single goal, there is likely to be a broad *horizon* of appropriate goals (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. In this context, it is worth noting that there are significant cultural differences in how to use this information. 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 round the class, and then will select some major issues for discussion with the whole class. This is consistent with what

Bromme and Steinbring (1994) discovered in their "expert-novice" analysis of two mathematics teachers: the novice teacher tended to treat students' questions as being from individual learners, while the expert teacher's responses tended to be directed more to a "collective student".

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

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

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

The second point is that the situation that the teacher set up in the first place—the question she chose to ask—has little potential for providing the teacher with useful information about the students' thinking, except, possibly, whether the students can recall the word "equilateral." This is typical in situations where the questions that the teacher uses in whole-class interaction have not been prepared in advance (in other words, when there is little or no pro-active or "upstream" regulation).

In contrast, the vignettes of mathematics teaching given in Kilpatrick *et al*. (2001) show how exemplary teachers actually design these "teachable moments" into their lessons: "Mr. Hernandez and Ms. Kaye have each designed the lesson to afford them critical information about their students' progress. The tasks they frame create a strategic space for students' work and for gaining insights into students' thinking" (p. 349). Similar considerations apply when the teacher collects in the students' notebooks and attempts to give helpful feedback to the students in the form of comments on how to improve rather than grades or percentage scores. If sufficient attention has not been given "upstream" to the design of the tasks given to the students, then the teacher may find that she has nothing useful to say to the students. Ideally, from examining the students' responses to the task, the teacher would be able to judge how to (a) help the learners learn better and (b) what she might do to improve the teaching of this topic. In this way, the assessment could be formative for the students, through the feedback she provides, and formative for the teacher herself, in that appropriate analysis of the students responses might suggest how the lesson could be improved.

As a concrete illustration of these ideas in action, consider the following account of a teacher participating in a research project (Lyon, Leahy, Morris & Thompson, 2005) who was working through a task entitled "Up, up and away" (Marquez & Boxley, 2003). The task requires students to complete a table and a graph for a weather balloon that rises at a rate of 8 feet per second. The students were then given a graph of a flare that was set off 2 seconds after the weather balloon and asked three questions:

1. What is the maximum height, in feet, that the flare reaches?
2. From the time the flare is set off, how many seconds will the flare take to reach its maximum height?
3. What is the average speed of the flare, in feet per second, from the ground to its highest point from the ground?

This particular class began to work in groups when the graph of the flare was introduced. The teacher said to the class, *"I'm hearing some problems, especially for the third part. Write your answers to the first part on the dry erase*." All the answers held up by the students were correct.

The teacher then asked the students to hold up their answers to the second question. She said, *"We have 3 different answers: 3, 4, and 5 seconds. It hits the maximum at what time?"* One student answered, *"5 seconds."*

The teacher asked, *"And starts at?"* A student said, *"2 seconds."*

The teacher asked the students to hold up their answers for the third question and saw that there were still different answers. She said, *"We have different answers. We are going to look at a different problem"*.

She then took the students through a similar activity, again looking at average speed, but this time in the context of car journeys where speed was measured in miles per hour. After some discussion, students showed, again by displaying their answers on dry-erase boards, that they understood how to calculate from a distance-time graph, the average speed over a specified time interval of a body moving at non-constant speed.

The students then resumed work on the "Up, up, and away" task. After a few moments the teacher asked the groups to hold up their new answers to the third question. Almost all the posted answers were correct, so the teacher asked *"If the average speed is 48 feet per second does that mean that the flare is always traveling at that speed?"* The students answered, *"To have a constant speed the graph would be a straight line*.

In the post-interview the teacher was asked if this was an on-the fly change to the lesson plan or if the additional problem had been part of her lesson design. The teacher had anticipated that the students might have a problem with the average speed and therefore designed the problem in case there was difficulty during the lesson. She only planned on using the additional problem if it were needed.

This brief extract illustrates several features of the effective regulation of learning. Before the lesson began, the teacher had planned the questions she was going to ask, and had also provided the students with dry erase boards, so that she could require responses from all students. She had also developed, within the students, a willingness to be open about their answers, as a step towards the "metacognitive culture" mentioned by Perrenoud above. These are examples of the "pro-active" regulation of learning. As a result of teaching this material previously, she had also identified that the third question of the three might pose particular difficulties—in other words, that this was a "hinge-point" in the lesson—and after some reflection on why this might have occurred, she planned the backup activity. This use of reflection on past teaching to modify future teaching is an example of "retro-active" regulation of learning. During the lesson, there was also a significant amount of the "interactive" regulation of learning. She listened carefully to the students' discussions, and looked at their work as she walked around the classroom. It was as a result of this that she

said, “I’m hearing some problems” and decided to collect evidence more systematically about the understanding of her students by asking them to display their responses on dry-erase boards. Because she had evidence from all students, she was in a much better position to make the “on-the-fly” decision to use the “backup” activity she had planned. It is also worth noting that the teacher used “all-student response systems” selectively. At times, she relied on her intuitive perceptions of the class, at times she decided to collect information systematically, and at other times, she allowed students to volunteer contributions. Shulman (2005) describes practices such as these as “pedagogies of uncertainty.” This term accurately reflects the fact that it is impossible to be certain about what students will learn as a result of a particular episode of instruction but the term “uncertainty” suggests that it is impossible to predict anything in advance. As the vignette above shows, in many cases, we might not know what will happen, but we can often reduce the uncertainty by careful planning. In this case, the teacher reduced the possibilities, to a simple decision about whether to use the supplementary activity she had prepared or not. In other cases, the teacher may plan for a more complex range of possibilities, as in the case with the item that asked students for a fraction between one-sixth and one-seventh, but again, careful consideration of the kinds of responses that students may produce allows teachers to “pre-filter” the students’ responses. By anticipating the students’ responses the teacher can simplify the task she is faced with in the classroom. Rather than “pedagogies of uncertainty,” therefore, it seems more appropriate to describe such pedagogies as “pedagogies of contingency.” The essence of formative assessment is that instruction is contingent on what it is that the students have learned. Teachers can regulate learning proactively by creating “moments of contingency,” for example by identifying a “hinge point” in a sequence of instruction as discussed above, and designing a “hinge-point” question to be used at that time. Teachers can also regulate learning interactively by capitalizing on moments of contingency that may arise spontaneously in their teaching. Through careful reflection after the lesson, they can plan better for the future, both in terms of how to create more “moments of contingency,” and in terms of how to recognize and take advantage of such moments when they arise.

Summary

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

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

be making frequent adjustments to the instruction to better meet the learning needs of the students.

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

References

Adhami, M., Johnson, D. C., & Shayer, M. (1997, November). *Does CAME work? Summary report on phase 2 of the Cognitive Acceleration in Mathematics Education (CAME) project*. Paper presented at the Day Conference of the British Society for Research into Learning Mathematics, Bristol, UK. Retrieved on May 14, 2006 from http://www.bsrlm.org.uk/IPs/ip17-3/BSRLM-IP-17-3-Full.pdf

Adhami, M., Johnson, D. C., & Shayer, M. (1998). *Thinking maths: accelerated learning in mathematics (KS3 and S1/S2)*. Oxford, UK: Heinemann Educational.

Ames, C. (1992). Classrooms: goals, structures, and student motivation. *Journal of Educational Psychology,* **84**(3), 261-271.

Antil, L. R., Jenkins, J. R., Wayne, S. K., & Vadasy, P. F. (1998). Cooperative learning: prevalence, conceptualization and the relation between research and practice. *American Educational Research Journal,* **35**(3), 419-454.

Arter, J. A., & McTighe, J. (2001). *Scoring rubrics in the classroom*. Thousand Oaks, CA: Corwin Press.

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

Bandura, A. (1977). Self-efficacy: towards aunifying theory of behavioral change. *Psychological Review,* **84**(2), 191-215.

Bandura, A. (1986). *Social foundations of thought and action: a social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1997). *Self-efficacy: the exercise of control*. New York, NY: W. H. Freeman.

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

Barnett, W. S., & Camilli, G. (2001). Compensatory preschool education, cognitive development and "race". In J. M. Fish (Ed.), *Race and intelligence: separating science from myth* (pp. 369-406). Mahwah, NJ: Lawrence Erlbaum Associates.

Bergan, J. R., Sladeczek, I. E., Schwarz, R. D., & Smith, A. N. (1991). Effects of a measurement and planning system on Kindergartners' cognitive development and educational programming. *American Educational Research Journal,* **28**(3), 683-714.

Berliner, D. C. (1994). Expertise: the wonder of exemplary performances. In J. N. Mangieri & C. C. Block (Eds.), *Creating powerful thinking in teachers and students: diverse perspectives* (pp. 161-186). Fort Worth, TX: Harcourt Brace College.

Bernstein, B. (Ed.). (1975). *Class, codes and control volume 3: towards a theory of educational transmissions* (Vol. 3). London, UK: Routledge and Kegan Paul.

Bertua, C., Anderson, N., & Salgado, J. F. (2006). The predictive validity of cognitive ability tests: a UK meta-analysis. *Journal of Occupational and Organisational Psychology*.

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

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

Black, P. 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)* (pp. 183-188). Chicago, IL: 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)* (Vol. Part II, pp. 20-50). Chicago, IL: 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* (pp. 81-100). London, UK: Sage.

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.

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.

Boaler, J. (2002). *Experiencing school mathematics: traditional and reform approaches to teaching and their impact on student Learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Boaler, J., & Humphreys, C. (2005). *Connecting mathematical ideas: middle school video cases to support teaching and learning*. Portmsouth, NH: Heinemann.

Boaler, J., Wiliam, D., & Zevenbergen, R. (2000). The construction of identity in secondary mathematics education. In J. F. Matos & M. Santos (Eds.), *Mathematics Education and Society* (pp. 192-202). Montechoro, Portugal: Centro de Investigação em Educação da Faculdade de Ciências Universidade de Lisboa.

Boekaerts, M. (1993). Being concerned with well being and with learning. *Educational Psychologist,* **28**(2), 149-167.

Boekaerts, M. (2001). Context sensitivity: activated motivational beliefs, current concerns and emotional arousal. In S. Volet & S. Järvelä (Eds.), *Motivation in learning contexts: theoretical advances and methodological implications* (pp. 17-31). Oxford, UK: Pergamon.

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.

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: a perspective on assessment and intervention. *Applied Psychology: An International Review,* **54**(2), 199-231.

Boekaerts, M., Maes, S., & Karoly, P. (2005). Self-regulation across domains of applied psychology: is there an emerging consensus? *Applied Psychology: An International Review,* **54**(2), 149-154.

Boykin, A. W., Coleman, S. T., Lilja, A., & Tyler, K. M. (2004). *Building on children's cultural assets in simulated classroom performance environments: research vistas in the communal learning paradigm* (Report no. 68). Washington, DC: Howard University.

Boykin, A. W., Lilja, A., & Tyler, K. M (2004). The influence of communal vs. individual learning context on the academic performance in social studies of grade 4–5 African-Americans. *Learning Environments Research,* **7**(3), 227-244.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: brain, mind, experience, and school*. Washington, DC: National Academy Press.

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.

Bromme, R., & Steinbring, H. (1994). Interactive development of subject matter in the mathematics classroom. *Educational Studies in Mathematics,* **27**(3), 217-248.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education,* **10**(2), 161-180.

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

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

Brousseau, G. (1984). The crucial role of the didactical contract in the analysis and construction of situations in teaching and learning mathematics (G. Seib, Trans.). In H.-G. Steiner (Ed.), *Theory of mathematics education: ICME 5 topic area and miniconference* (Vol. 54, pp. 110-119). Bielefeld, Germany: Institut für Didaktik der Mathematik der Universität Bielefeld.

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

Brown, A. (1987). Metacognition, executive control, self-regulation and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation and understanding* (pp. 65-116). HIllsdale, NJ: Lawrence Erlbaum Associates.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: a theoretical synthesis. *Review of Educational Research,* **65**(3), 245-281.

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

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

Butler, R., & Nisan, M. (1986). Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology,* **78**(3), 210-216.

Cameron, J. A., & Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation: a meta-analysis. *Review of Educational Research,* **64**(3), 363-423.

Carol. (2006, Feb 22). *Conversations with children and literacy*. Retrieved 19 May, 2006, from feed://themediansib.com/category/this-n-that/feed/

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C. P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: an experimental study. *American Educational Research Journal,* **26**(4), 499-531.

Carr, M. M., Alexander, J., & Folds-Bennett, T. (1994). Metacognition and mathematics strategy use. *Applied Cognitive Psychology,* **8**, 583-595.

Ciofalo, J., & Wylie, E.C. (2006). Using diagnostic classroom assessment: one question at a time. *Teachers College Record,* Date published: 10 Jan 2006 http://www.tcrecord.org/content.asp?contentid=12285, Date accessed 18 Jan 2006.

Clarke, D. (1996). Assessment. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 327-370). Dordrecht, Netherlands: Kluwer Academic Publishers.

Clarke, S. (2001). *Unlocking formative assessment*. London, UK: Hodder & Stoughton.

Claxton, G. L. (1995). What kind of learning does self-assessment drive? Developing a 'nose' for quality: comments on Klenowski. *Assessment in Education: principles, policy and practice,* **2**(3), 339-343.

Cohen, E. G. (1994). Restructuring the classroom: conditions for productive small groups. *Review of Educational Research,* **64**(1), 1-35.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: a motivational analysis of self-system processes. In M. R. Gunnar & L. A. Sroufe (Eds.), *Minnesota symposia on child psychology* (Vol. 23, pp. 43-77). Hillsdale, NJ: Lawrence Erlbaum Associates.

Corno, L. (2001). Volitional aspects of self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated leaning and academic achievement: theoretical perspectives* (2 ed., pp. 191-225). Hillsdale, NJ: Lawrence Erlbaum Associates.

Covington, M. (1992). *Making the grade: a self-worth perspective on motivation and school reform*. New York, NY: Cambridge University Press.

Crespo, S. (2000). Seeing more than right and wrong answers: prospective teachers' interpretations of students' mathematical work. *Journal of Mathematics Teacher Education,* **3**, 155-181.

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

Csikszentmihalyi, M. (1990). *Flow: the psychology of optimal experience*. New York, NY: Harper & Row.

Davis, B. (1997). Listening for differences: an evolving conception of mathematics teaching. *Journal for Research in Mathematics Education,* **28**(3), 355-376.

Dawes, L., Mercer, N., & Wegerif, R. (2000). *Thinking Together: a programme of activities for developing thinking skills at KS2*. Birmingham, UK: Questions Publishing Company.

Day, J. D., & Cordón, L. A. (1993). Static and dynamic measures of ability: an experimental comparison. *Journal of Educational Psychology,* **85**(1), 76-82.

Deary, I. (2000). *Looking down on human intelligence* (Vol. 34). Oxford, UK: Oxford University Press.

Deci, E. L. (1996). *Why we do what we do*. New York, NY: Penguin.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum.

Deci, E. L., & Ryan, R. M. (1994). Promoting self-determined education. *Scandinavian Journal of Educational Research,* **38**(1), 3-14.

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

DeCorte, E. & Verschaffel, L. (2006). Mathematical thinking and learning. In K. A. Renninger & I. E. Sigel (Eds.), *Hanbook of child psychology volume 4: child psychology in practice* (6 ed., pp. 103-152). New York, NY: Wiley.

DeCorte, E., Verschaffel, L., & Op't Eynde, P. (2000). Self-regulation: a characteristic and a goal of mathematics education. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 687-726). San Diego, CA: Academic Press.

Deevers, M. (2006). *Linking classroom assessment practices with student motivation in mathematics* Paper presented at Annual meeting of the American Educational Research Association held at San Francisco, CA. Cleveland, OH: Cleveland State University.

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.

Dewey, J. (1913). *Interest and effort in education*. Boston, MA: Riverside Press.

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

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

diSessa, A. A., & Minstrell, J. (1998). Cultivating conceptual change with benchmark lessons. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematical and science learning* (pp. 155-187). Mahwah, NJ: Lawrence Erlbaum Associates.

Donovan, M. S., & Bransford, J. (Eds.). (2005). *How students learn: history, mathematics and science in the classroom*. Washington, DC: National Academies Press.

Dorr-Bremme, D. W., & Herman, J. L. (1986). *Assessing student achievement: a profile of classroom practices* (Vol. 11). Los Angeles, CA: University of California Los Angeles Center for the Study of Evaluation.

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

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

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75-146). San Francisco: W H Freeman.

Eisenberger, R., & Cameron, J. A. (1996). Detrimental effects of reward: reality or myth? *American Psychologist,* **51**(11), 1153-1166.

Elawar, M. C., & Corno, L. (1985). A Factorial experiment in teachers' written feedback on student homework: changing teacher behavior a little rather than a lot. *Journal of Educational Psychology,* **77**(2), 162-173.

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

English, L. D., & Doerr, H. M. (2004). *Listening and responding to students' ways of thinking*. Paper presented at the 27th annual conference of the Mathematics Education Research Group of Australasia: Mathematics education for the third millennium: towards 2010, held at Townsville, Queensland, Australia. Sydney, Australia: MERGA Inc.

Ernest, P. (1991). *The philosophy of mathematics education* (Vol. 1). London: Falmer Press.

Even, R., & Tirosh, D. (1995). Subject-matter knowledge and the knowledge about students as sources of teacher presentations of the subject-matter. *Educational Studies in Mathematics,* **29**(1), 1-20.

Even, R., & Tirosh, D. (2002). Teacher knowledge and understanding of students' mathematical learning. In L. D. English (Ed.), *Handbook of international research in mathematics education* (pp. 219-240). Mahwah, NJ: Lawrence Erlbaum Associates.

Fantuzzo, J. W., Davis, G. Y., & Ginsburg, M. D. (1995). Effects of parent involvement in isolation or in combination with peer tutoring on student self-concept and mathematics achievement. *Journal of Educational Psychology,* **87**(2), 272-281.

Fantuzzo, J. W., King, J., & Heller, L. R. (1992). Effects of reciprocal peer tutoring on mathematics and school adjustment: a component analysis. *Journal of Educational Psychology,* **84**, 331-339.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education,* **27**(4), 403-434.

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

Fish, J. M. (Ed.). (2001). *Race and intelligence: separating science from myth*. Mahwah, NJ: Lawrence Erlbaum Associates.

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 231-235). Hillsdale, NJ: Lawrence Erlbaum Associates.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. *American Psychologist,* **34**, 906-911.

Flavell, J. H. (1981). Cognitive monitoring. In W. P. Dickson (Ed.), *Children's oral communication skills* (pp. 35-60). New York, NY: Academic Press.

Flynn, J. R. (1984). The mean IQ of Americans: massive gains 1932 to 1978. *Psychological Bulletin,* **95**(1), 58-64.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: what IQ tests really measure. *Psychological bulletin,* **101**(2), 171-191.

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

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

Franke, M. L., Carpenter, T. P., Levi, L., & Fennema, E. (2001). Capturing teachers' generative change: a follow-up study of professional development in mathematics. *American Educational Research Journal,* **38**(3), 653-689.

Franke, M. L., Fennema, E., & Carpenter, T. P. (1997). Teachers creating change: examining evolving beliefs and classroom practices. In E. Fennema & B. S. Nelson (Eds.), *Mathematics teachers in transition* (pp. 255-282). Mahwah, NJ: Lawrence Erlbaum Associates.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal,* **28**(3), 617-641.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Dutka, S., & Katzaroff, M. (1996). The relation between student ability and the quality and effectiveness of explanations. *American Educational Research Journal,* **33**(3), 631-664.

Fuson, K. C., Kalchman, M., & Bransford, J. D. (2005). Mathematical understanding: an introduction. In M. S. Donovan & J. Bransford (Eds.), *How students learn: history, mathematics and science in the classroom* (pp. 217-256). Washington, DC: National Academies Press.

Gay, S., & Thomas, M. (1993). Just because they got it right, does it mean they know it? In N. L. Webb & A. F. Coxford (Eds.), *Assessment in the mathematics classroom: 1993 yearbook of the National Council of Teachers of Mathematics* (pp. 130-134). Reston, VA: National Council of Teachers of Mathematics.

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

Girling, M. (1977). Towards a definition of basic numeracy. *Mathematics teaching*(81), 4-5.

Glaser, R., & Silver, E. A. (1994). Assessment, testing and instruction: retrospect and prospect. In L. Darling-Hammond (Ed.), *Review of research in education* (Vol. 20, pp. 393-419). Washington, DC: American Educational Research Association.

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

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

Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: an experiemental and individual difference investigation. *Journal of Personality and Social Psychology,* **52**(5), 890-898.

Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.). (1998). *Metacognition in educational theory and practice*. Mahwah, NJ: Lawrence Erlbaum Associates.

Harlen, W., & Deakin-Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning (version 1.1). *Research Evidence in Education Library*. London, UK: University of London Institute of Education Social Science Research Unit. Retrieved on May 28th, 2006 from http://eppi.ioe.ac.uk/EPPIWebContent/reel/review_groups/assessment/ass_rv1/ass_rv1.pdf

Hart, K. M. (Ed.). (1981). *Children's understanding of mathematics: 11-16*. London, UK: John Murray.

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

Heid, M. K., Blume, G. W., Zbiek, R. M., & Edwards, B. S. (1999). Factors that influence teachers learning to do interviews to understand students' mathematical understandings. *Educational Studies in Mathematics,* **37**(3), 223-249.

Helmke, A., & Schrader, F. W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education,* **3**(2), 91-98.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve*. New York, NY: Free Press.

Hickey, D. T., & McCaslin, M. (2001). A comparative, sociocultural analysis of context and motivation. In S. Volet & S. Järvelä (Eds.), *Motivation in learning contexts: theoretical advances and methodological implications* (pp. 33-55). Oxford, UK: Pergamon.

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically umotivated: a critical issue for the 21st century. *Review of Educational Research,* **70**(2), 151-179.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J. K., Miu-Ying Chui, A., Wearne, D., Smith, M., Kersting, N., Manaster, A., Tseng, E., Etterbeek, W., Manaster, C., Gonzales, P., & Stigler, J. W. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 Video Study* (Vol. NCES (2003–013)). Washington, DC: National Center for Education Statistics.

Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., Hollingsworth, H., Manaster, A., Wearne, D., & Gallimore, R. (2005). Mathematics teaching in the United States today (and tomorrow): results from the TIMSS 1999 Video Study. *Educational Evaluation and Policy Analysis,* **27**(2), 111-132.

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

Hoek, D., van den Eeden, P., & Terwel, J. (1999). The effects of integrated social and cognitive strategy instruction on the mathematics achievement in secondary education. *Learning and Instruction,* **9**(5), 427-448.

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

Järvelä, S. (2001). Shifting research on motivation and cognition to an integrated approach on learning and motivation in context. In S. Volet & S. Järvelä (Eds.), *Motivation in learning contexts: theoretical advances and methodological implications* (pp. 3-14). Oxford, UK: Pergamon.

Keddie, N. (1971). Classroom knowledge. In M. F. D. Young (Ed.), *Knowledge and control*. London, UK: Collier-Macmillan.

Kilpatrick, J., Swafford, J. O., & Findell, B. (Eds.). (2001). *Adding it up: helping children learn mathematics*. Washington, DC: National Academy Press.

Klein, D. (2003). A brief history of American K-12 mathematics education in the 20th century. In J. M. Royer (Ed.), *Mathematical Cognition*. Greenwich, CT: Information Age Publishing.

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.

Kluwe, R. H. (1982). Cognitive knowledge and executive control: metacognition. In D. R. Griffin (Ed.), *Animal mind—human mind* (pp. 201-224). New York, NY: Springer-Verlag.

Kohn, A. (1999). *Punished by rewards: the trouble with gold stars, incentive plans, A's, praise and other bribes* (2 ed.). Boston, MA: Houghton-Mifflin.

Kohn, A. (2006). The trouble with rubrics. *English Journal,* **95**(4), 12-15.

Kramarski, B., Mevarech, Z. R., & Arami, M. (2002). The effects of metacognitive instruction on solving mathematical authentic tasks. *Educational Studies in Mathematics,* **49**(2), 225-250.

Kramarski, B., Mevarech, Z. R., & Lieberman, A. (2001). Effects of multilevel versus unilevel metacognitive training on mathematical reasoning. Journal of Educational Research, 94, 292-300. *Journal of Educational Research,* **94**, 292-300.

Lampert, M. (2001). *Teaching problems and the problems of teaching*. New Haven, CT: Yale University Press.

Lave, J., Murtaugh, M., & de la Roche, O. (1984). The dialectic of arithmetic in grocery shopping. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: its development in social context* (pp. 67-94). Cambridge, MA: Harvard University Press.

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

Lesh, R., Hoover, M., Hole, B., Kelly, A. E., & Post, T. (2001). Principles for developing thought-revealing activities for students and teachers. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 591-646). Mahwah, NJ: Lawrence Erlbaum Associates.

Lester Jr, F. K., Lambdin, D. V., & Preston, R. V. (1997). A new vision of the nature and purposes of assessment in the mathematics classroom. In G. D. Phye (Ed.), *Handbook of classroom assessment: learning, adjustment and achievement* (pp. 287-319). San Diego, CA: Academic Press.

Linnenbrink, E. A. (2005). The dilemma of performance-approach goals: the use of multiple goal contexts to promote students' motivation and learning. *Journal of Educational Psychology,* **97**(2), 197–213.

Lodico, M. G., Ghatala, E. S., Levin, J. R., Pressley, M., & Bell, J. A. (1983). The effects of strategy-monitoring training on children's selection of effective memory strategies. *Journal of Experimental Child Psychology,* **35**(2), 263-277.

Looney, J. (Ed.). (2005). *Formative assessment: improving learning in secondary classrooms*. Paris, France: Organisation for Economic Cooperation and Development.

Lyon, C., Leahy, S., Morris, T., & Thompson, M. (2005). *Lessons learned (the hard way) in the evidence-centered teaching in algebra project*. Research Memorandum. Princeton, NJ: Educational Testing Service.

Mackintosh, N. J. (2000). *IQ and human intelligence*. Oxford, UK: Oxford University Press.

Marquez, E., & Boxley, B. (2003). *Teacher assistance package guide 3: using mathematical models to represent and understand quantitative relationships*. Princeton, NJ: Educational Testing Service.

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.

Marso, R. N., & Pigge, F. L. (1993). Teachers' testing knowledge, skills, and practices. In S. L. Wise (Ed.), *Teacher training in measurement and assessment skills* (pp. 129-185). Lincoln, NE: Buros Institute of Mental Measurements.

Mayer, D. K., & Turner, J. C. (2002). Using instructional discourse analysis to study the scaffolding of student self-regulation. *Educational Psychologist,* **37**(1), 17-25.

Mayer, D. K., Turner, J. C., & Spencer, C. A. (1997). Challenge in a mathematics classroom: students' motivation and strategies in project-based learning. *Elementary School Journal,* **97**(5), 501-521.

McCaslin, M., & Good, T. L. (1996). The informal curriculum. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 622-670). New York, NY: Macmillan.

McCaslin, M., & Hickey, D. T. (2001). Educational psychology, social constructivism, and educational practice: a case of emergent identity. *Educational Psychologist,* **36**(2), 133-140.

McClain, K., & Cobb, P. (2001). An analysis of development of sociomathematical norms in one first-grade classroom. *Journal for Research in Mathematics Education,* **32**(3), 236-266.

McMorris, R. F., & Boothroyd, R. A. (1993). Tests that teachers build: an analysis of classroom tests in science and mathematics. *Applied Measurement in Education,* **6**, 321-342.

Mehan, H. (1979). *Learning lessons: social organization in the classroom*. Cambridge, MA: Harvard University Press.

Meisels, S. J., Atkins-Burnett, S., Xue, Y., Bickel, D. D., & Son, S.-H. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. *Education Policy Analysis Archives,* **11**(9). Retrieved 12.12.05 from http://epaa.asu.edu/epaa/v11n9/

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.

Mevarech, Z. R., & Kramarski, B. (1997). IMPROVE: a multidimensional method for teaching mathematics in heterogeneous classrooms. *American Educational Research Journal,* **34**(2), 365-394.

Mevarech, Z. R., & Kramarski, B. (2003). The effects of metacognitive training versus worked-out examples on students' mathematical reasoning. *British Journal of Educational Psychology,* **73**(4), 449-471.

Mitchell, M. (1993). Situational interest: its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology,* **85**, 424-436.

Montague, A. (Ed.). (1999). *Race and IQ*. New York, NY: Oxford University Press.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

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

Neisser, U. (Ed.). (1998). *The rising curve: long-term gains in IQ and related measures*. Washington, DC: American Psychological Association.

Nisbett, R. E., & Ross, L. D. (1980). *Human inference: strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

Niss, M. (1993). Assessment in mathematics education and its effects. In M. Niss (Ed.), *Investigations into assessment in mathematics education: an ICMI study* (pp. 1-30). Dordrecht, Netherlands: Kluwer Academic Publishers.

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.

Op't Eynde, P., DeCorte, E., & Verschaffel, L. (2001). "What to learn from what we feel?" The role of students' emotions in the mathematics classroom. In S. Volet & S. Järvelä (Eds.), *Motivation in learning contexts: theoretical advances and methodological implications* (pp. 149-167). Oxford, UK: Pergamon.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: the science and design of educational assessment*. Washington, DC: National Academy Press.

Perrenoud, P. (1991). Towards a pragmatic approach to formative evaluation. In P. Weston (Ed.), *Assessment of pupil achievement* (Vol. Part A: 25, 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.

Plomin, R., & Petrill, S. A. (1997). Genetics and intelligence: what's new? *Intelligence,* **24**, 53-77.

Pryor, J., & Crossouard, B. (2005). *A sociocultural theorization of formative assessment* Paper presented at Sociocultural Theory in Educational Research and Practice Conference held at University of Manchester.Brighton, UK: University of Sussex

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

Raven, J. (1960). *Guide to the standard progressive matrices: sets A, B, C and D*. London, UK: Lewis.

Relearning by Design. (2000). What is a rubric? Retrieved on May 1st, 2006 from http://www.relearning.org/resources/PDF/rubric_sampler.pdf.

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

Romberg, T. A. (1992). Perspectives on scholarship and research methods. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 49-64). New York, NY: Macmillan.

Roschelle, J., Abrahamson, L., & Penuel, W. R. (2004). *Integrating classroom network technology and learning theory to improve classroom science learning: a literature synthesis*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA. Menlo Park, CA: SRI International.

Rosenshine, B. V., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: a review of intervention studies. *Review of Educational Research,* **66**(2), 181-221.

Ross, J. A. (1995). Effects of feedback on student behavior in cooperative learning groups in a grade-7 math class. *Elementary School Journal,* **96**(2), 125-143.

Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5-6 mathematics: effects on problem solving achievement. *Educational Assessment,* **8**(1), 43-58.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (2002). Influences on student cognitions about evaluation. *Assessment in Education: Principles Policy and Practice,* **9**(1), 81-95.

Rowan, B., Harrison, D. M., & Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *Elementary School Journal,* **105**, 103-127.

Rowe, M. B. (1974). Wait time and rewards as instructional variables, their influence on language, learning and fate control. *Journal of Research in Science Teaching,* **11**, 81-94.

Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: examining reasons for acting in two domains. *Journal of Personality and Social Psychology,* **57**, 749-761.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemporary Educational Psychology,* **25**, 54-67.

Ryan, R. M., & Stiller, J. (1991). The social contexts of internalization: parent and teacher influences on autonomy, motivation, and learning. In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 7, pp. 115-149). Greenwich, CT: JAI Press.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education,* **13**, 191-209.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science,* **18**, 119-144.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: what's really happening? *Phi Delta Kappan,* **62**(5), 631-634.

Saphier, J. (2005). Masters of motivation. In R. DuFour, R. Eaker & R. DuFour (Eds.), *On common ground: the power of professional learning communities* (pp. 85-113). Bloomington, IL: National Education Service.

Schoenfeld, A. H. (1985). *Mathematical problem-solving*. New York, NY: Academic Press.

Schoenfeld, A. H. (1992). Learning to think mathematically: problem solving, metacognition and sense making in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334-370). New York, NY: Macmillan.

Schunk, D. H. (1990). Goal-setting and self-efficacy during self-regulated learning. *Educational Psychologist,* **25**, 71-86.

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist,* **26**, 207-231.

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.

Selden, S. (1999). *Inheriting shame: the story of eugenics and racism in America*. New York, NY: Teachers College Press.

Senk, S. L., Beckman, C. E., & Thompson, D. R. (1997). Assessment and grading in high school mathematics classrooms. *Journal for Research in Mathematics Education,* **28**(2), 187-215.

Sfard, A. (1998). On two metaphors for learning and on the dangers of choosing just one. *Educational Researcher,* **27**(2), 4-13.

Shacter, J. (2000). Does individual tutoring produce optimal learning? *American Educational Research Journal,* **37**(3), 801-829.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: political rhetoric and measurement reality. *Educational Researcher,* **21**(4), 22-27.

Shavelson, R. J., Black, P. J., Wiliam, D., & Coffey, J. (2003). *On aligning formative and summative functions in the design of large-scale assessment systems*. Paper presented at National Research Council workshop on Assessment In Support of Instruction and Learning: Bridging the Gap Between Large-Scale and Classroom Assessment. Stanford, CA: Stanford University School of Education.

Shayer, M. (2003). Not just Piaget; not just Vygotsky, and certainly not Vygotsky as alternative to Piaget. *Learning and Instruction,* **13**(5), 465-485.

Shepard, L. A., Hammerness, K., Darling-Hammond, L., Rust, F., Snowden, J. B., Gordon, E., Gutierrez, C., & Pacheco, A. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: what teachers should learn and be able to do* (pp. 275-326). San Francisco, CA: Jossey-Bass.

Shulman, L. S. (2005). *The signature pedagogies of the professions of law, medicine, engineering, and the clergy: potential lessons for the education of teachers* Paper presented at National Science Foundation Mathematics and Science Partnerships Workshop: Teacher Education for Effective Teaching and Learning held at National Research Council Center for Education, Irvine, CA.

Siero, F., & van Oudenhoven, J. P. (1995). The Effects of Contingent Feedback on Perceived Control and Performance. *European Journal of Psychology of Education,* **10**(1), 13-24.

Simmons, M., & Cope, P. (1993). Angle and rotation: effects of differing types of feedback on the quality of response. *Educational Studies in Mathematics,* **24**(2), 163-176.

Skemp, R. R. (1977). Relational understanding and instrumental understanding. *Mathematics teaching*(77), 20-26.

Slavin, R. E. (1995). *Cooperative learning: theory, research and practice* (2 ed.). Boston, MA: Allyn & Bacon.

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

Smith, E., & Gorard, S. (2005). 'They don't give us our marks': the role of formative feedback in student progress. *Assessment in Education: Principles Policy and Practice,* **12**(1), 21-28.

Sorrentino, R. M., & Higgins, E. T. (1986). Motivation and synergism: warming up to synergism. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: foundations of social behavior* (pp. 3-19). New York, NY: Guildford Press.

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering and technology: a meta-analysis. *Review of Educational Research,* **69**(1), 21-51.

Sternberg, R. J., & Williams, W. (1998). Applying the triarchic theory of human intelligence in the classroom. In R. J. Sternberg & W. Williams (Eds.), *Intelligence, instruction and assessment: theory into practice* (pp. 1-15). Mahwah, NJ: Lawrence Erlbaum Associates.

Stevens, R. J., & Slavin, R. E. (1995). Effects of a cooperative learning approach in reading and writing on academically handicapped and nonhandicapped students. *Elementary School Journal,* **95**(3), 241-262.

Stiggins, R. J. (2001). *Student-involved classroom assessment* (3 ed.). Upper Saddle River, NJ: Prentice-Hall.

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

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

Stillman, G. A., & Galbraith, P. L. (1998). Applying mathematics with real world connections: metacognitive characteristics of secondary students. *Educational Studies in Mathematics,* **36**(2), 157-194.

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

Tobin, K. (1987). The role of wait time in higher cognitive level learning. *Review of Educational Research,* **57**(1), 69-95.

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

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment*. Buckingham, UK: Open University Press.

Tunstall, P., & Gipps, C. V. (1996a). Teacher feedback to young children in formative assessment: a typology. *British Educational Research Journal,* **22**(4), 389-404.

Tunstall, P., & Gipps, C. V. (1996b). 'How does your teacher help you to make your work better ?' Children's understanding of formative assessment. *The Curriculum Journal,* **7**(2), 185-203.

van den Heuvel-Panhuizen, M., & Becker, J. (2003). Towards a didactic model for assessment design in mathematics education. In A. Bishop, M. A. Clements, C. Keitel, J.

Kilpatrick & F. K. S. Leung (Eds.), *Second International Handbook of Mathematics Education* (pp. 689-716). Dordrecht, Netherlands: Kluwer Academic Publishers.

VanLehn, K. (1990). *Mind bugs: the origins of procedural misconceptions*. Cambridge, MA: MIT Press.

Vermunt, J. D. (2003). The power of learning environments and the quality of student learning. In E. DeCorte, L. Verschaffel, N. Entwistle & J. van Merriënboer (Eds.), *Powerful learning environments: unravelling basic components and dimensions* (pp. 109-124). Oxford, UK: Pergamon.

Vinner, S. (1997). From intuition to inhibition—mathematics, education and other endangered species. In E. Pehkonen (Ed.), *Proceedings of the 21st Conference of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 63-78). Lahti, Finland: University of Helsinki Lahti Research and Training Centre.

von Glasersfeld, E. (1987). Learning as a constructive activity. In C. Janvier (Ed.), *Problems of representation in the teaching and learning of mathematics* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates.

von Glasersfeld, E. (Ed.). (1991). *Radical constructivism in mathematics education*. Dordrecht, Netherlands: Kluwer.

Vygotsky, L. (1978). *The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Webb, D. C. (2004). Enriching classroom assessment opportunities through discourse. In T. A. Romberg (Ed.), *Standards-based mathematics assessment in middle school* (pp. 168-187). New York, NY: Teachers College Press.

Webb, N. L. (1992). Assessment of students' knowledge of mathematics: steps towards a theory. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 661-683). New York, NY: Macmillan.

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education,* **22**(5), 366-389.

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). *Looking inside the classroom: a study of K-12 mathematics and science education in the United States*. Chapel Hill, NC: Horizon Research Inc.

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

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

Wiener, N. (1948). *Cybernetics, or the control and communication in the animal and the machine*. New York, NY: John Wiley.

Wigfield, A., Eccles, J. S., & Rodriguez, D. (1998). The development of children's motivation in school contexts. In P. D. Pearson & A. Iran-Nejad (Eds.), *Review of research in education* (Vol. 23, pp. 73-118). Washington, DC: American Educational Research Association.

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.; 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. (to appear, 2007). Integrating assessment with instruction: what will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Williams, R. (1961). *The long revolution*. London, UK: Chatto & Windus.

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences,* **8**, 327-353.

Winne, P. H. (2005). Key issues in modeling and applying research on self-regulated learning. *Applied Psychology: An International Review,* **54**(2), 232-238.

Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education,* **27**(4), 458-477.

Zimmerman, B. J. (1986). Becoming a self-regulated learner: which are the key subprocesses? *Contemporary Educational Psychology,* **11**, 307-313.

Endnotes

---

[i] At various times in the preparation of this chapter, I received helpful comments from Susan M Brookhart, William S Bush, Edith Aurora Graf, Siobhan Leahy and Marnie Thompson, and as a result, the chapter changed considerably from the drafts on which they had commented. Consequently, I am solely responsible for the final version and any errors or omissions that remain, or indeed, were re-introduced.

[ii] In the USA, 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 by Mary James at an ASCD conference in New Orleans (James, 1992). Three years later, the phrase was used by Ruth Sutton as the title of a book (Sutton, 1995). However, the first use of the "prepositional permutation" appears to be the third edition of a book entitled "Assessment: a teacher's guide to the issues" by Caroline Gipps and Gordon Stobart, where the first chapter is entitled "Assessment of learning" and the second "Assessment for learning" (Gipps and Stobart, 1997). The distinction was brought to a wider audience by the Assessment Reform Group in 1989 in a guide for policy-makers (Broadfoot, Daugherty, Gardner, Gipps, Harlen, James & Stobart, 1989).

[iii] Here I am following the convention in American and British English that the term "assessment" applies to individuals, while "evaluation" applies to institutions or programs.

[iv] As Chaiklin (2003) makes clear, in reading Vygotsky, it is important to understand the cultural and historical context of his work, and in particular, Vygotsky's attempt to distinguish between learning and development. For Vygotsky, development requires changes in the psychological functions that the child can deploy, while learning does not. The zone of proximal development (ZPD) is not, therefore, just a way of describing what a student can do with support—that could just be learning. Rather the ZPD is a description of the *maturing* psychological functions rather than those that already exist, and a focus in instruction on the maturing psychological functions is most likely to produce a transition to the next developmental level.

[v] In English, the noun "regulation" has two meanings; one refers to the act of regulating and the other to a rule or law to govern conduct, and so, while it is the former sense that is intended here, the word has the unfortunate connotation of the second. In French, the two senses have separate terms (régulation and règlement) and so the problem does not arise.