Assessment and the regulation of learning[1][2]

Dylan Wiliam, ETS[3]

Introduction

Educational assessments are conducted in a variety of ways and their outcomes can be used for a variety of purposes. There are differences in who decides what is to be assessed, who carries out the assessment, where the assessment takes place, how the resulting responses made by students are scored and interpreted, and what happens as a result. In particular, each of these can be the responsibility of those who teach the students, while 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, education assessments serve three functions:

formative   supporting learning
summative   certifying individuals
evaluative   holding educational institutions accountable

Through a series of historical contingencies, we have arrived at a situation in the United States 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 all assessments with which we can hold educational institutions to account must be conducted by an external agency, even though in other countries, this is not the case.

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 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). 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. Many authors have argued that these tensions require that one assessment system cannot serve all functions and that separate systems are required. No matter how convincing the argument in favor of this suggestion, it seems to me that it must not be believed because the consequences are so deleterious for learning. Separate assessment systems result either in the exclusion of teachers from summative assessments, or requiring them to operate parallel but distinct assessment systems for summative and formative functions, which almost always results the marginalization of the formative function. If we are to develop integrated systems that can serve formative, summative and evaluative systems, the question that then arises is which functions should serve as the basis of the assessment. The position adopted in this paper is that the formative function should come first (Black and Wiliam, 2004b). The main reason for this is that fine-scale data that have been collected to support learning can always be aggregated to provide information on students and on institutions, but aggregated summative data on students and institutions cannot be disaggregated to serve learning needs. Tensions in the different uses of the data will, of course, remain, but these can be ameliorated, even if they can't be entirely eradicated.

Elsewhere, I have sketched out how an assessment system might be designed to serve all three functions reasonably well (Newton, 2003; Wiliam, 2003). In this paper, I want to outline a framework for thinking about the formative aspects of such a system, and in particular how assessment can be used to 'keep learning on track'.

What do we want formative assessment to do?

There is no agreed definition of the term 'formative assessment'. In the United States, the term is often used to describe assessments that are used to provide information on the likely performance of students on state-mandated tests—a usage than might better be described as 'early-warning summative'. In other contexts it is used to describe any feedback given to students, no matter what use it made of it, such as telling students which items they got correct and incorrect (sometimes called 'knowledge of results'). These kinds of usages suggest that the distinction between 'formative' and 'summative' applies to the assessments themselves, but since the same assessment can be used both formatively and summatively, it follows that these terms cannot describe assessment themselves, but the use to which the resulting outcomes are put.

In some contexts, assessments that are used to support learning are described under the broad heading 'assessment for learning' (in contrast to 'assessment *of* learning'). This does suggest a process, rather than being a description of the nature of the assessment itself, but the danger here is that the focus is placed on the intention behind the use of the assessment, rather than action that actually takes place (Wiliam and Black, 1996). Many writers use the terms 'assessment for learning' and 'formative assessment' interchangeably, but Black et al (2002, p. i) 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 paper, then, formative assessment refers 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). For this to happen, Ramaprasad (1983) suggests that we need four things:

• information about the current state of the system
• information about the desired state of the system
• a way to determining whether there is a 'gap' between these two
• a mechanism whereby the feedback can be used to 'close the gap' between the current state and the goal state.

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 was *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 change the quality of the learning within the system.

In order for assessment to function formatively at the classroom level, it needs to identify where learners are in their learning, where they are going, and how to get there. Crossing this three-fold typology of information needs with the different agents in the classroom (the student, her or his peers, and the teacher) creates the framework for looking at the role of formative assessment shown in figure 1.

	Teacher	Peer	Learner
Where the learner is	Evoking information	Peer-assessment	Self-assessment
Where she or he is going	Curriculum philosophy	Sharing success criteria	Sharing success criteria
How to get there	Feedback	Peer-tutoring	Self-directed learning

Figure 1: Aspects of formative assessment

To establish where the learner is in their learning, the teacher needs a range of ways of evoking information and eliciting the models that students hold (Lesh et al, 2003). This can be through questions, other prompts (including statements to which the students have to react), or through problem situations which reveal the schemas with which the students are operating. For example, after showing students a spring balance suspending a weight inside a bell jar, they can be asked what will happen to the reading on the spring balance if the air inside the bell jar is evacuated, and to explain their reasoning. Many students respond that the weight will rise, because the air is no longer pressing down on it, suggesting that they believe that objects have weight only because of the air-pressure acting downwards on the object (a belief no doubt reinforced by teachers' constant reminder of the 15 lbs per square inch of air pressure pressing in on each of us, and of film showing astronauts weighing less on the airless moon than on earth).

However, not all model-eliciting activities are equally important, even if they do tell us something new about students' conceptions. The choice of which models to elicit must be driven by a clear philosophy of the subject. For a given curriculum, some things are important to know and some are not and so it is also necessary to be clear about the desired outcome of the learning. In some cases, this may be a specific goal (eg getting the students to be able to find the area of a trapezoid, or balance a chemical equation) but in the case of many aspects of the language arts and social studies, as well as in open-ended and exploratory work in mathematics and science, there may be a whole range of goals that are appropriate for different learners or for learners at different stages of development.

Such prompts can also, as well as telling us where students are in their learning, can actually produce learning. For example, students who have become familiar with the notion of heat energy might be asked to estimate the heating requirements of a swimming pool of given volume. Such a task guides students towards the invention for themselves of the notion of specific heat capacity of water (i.e. the amount of energy needed to heat a kilogram of water by one degree).

Once the learning outcomes are clear, the provision of feedback from the teacher can assist learning, provided, of course, such feedback is acted upon. Several conditions need to be met for this to take place. The feedback itself needs to be task-involving rather than ego-involving (Kluger & DeNisi, 1996), but it also helps if students see the purpose of feedback as helping them improve, rather than simply judging their worth, if the students have mastery, rather than performance goals, and see ability as incremental rather than fixed (Dweck, 2000). Too often. as Perrenoud (1998) notes, "…feedback given to pupils in class is like so many bottles thrown out to the sea. No one can be sure that the message they contain will one day find a receiver" (p87).

Learning is also enhanced when learners are able to assess their own performance (Fontana & Fernandes, 1994). But as Sadler (1989) notes, this requires that learners come to understand the criteria for success that the teacher already has in mind. Learners often find this difficult, however, and the involvement of peers can help learners understand success criteria and monitor their own progress towards their goals (White & Frederiksen, 1998). Thus peer-assessment provides an important complement to, and may

even be a pre-requisite for, effective self-assessment (Black, Harrison, Lee, Marshall, & Wiliam, 2003).

Although we started out with a relatively simple idea—feedback—the formulation above presents rather a complex picture, and the ways in which the elements within figure 1 relate to each other is not straightforward. However, all the elements in figure 1 can be integrated within a more general theoretical framework of the *regulation of learning processes* as suggested Perrenoud (1991, 1998)[4]. 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.

Formative assessment and the regulation of learning

The first thing to say here is that it is important to distinguish between the regulation of the activity in which the student engages and the regulation of the learning that results. Most teachers appear to be quite skilled at the former, but have only a hazy idea of the learning that results. 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 from regulating the activities 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.

The second point to make is that the 'engineering of learning environments' does not guarantee that the learning is well-regulated. Many visual arts classroom are *productive*, in that they do lead to significant learning on the part of students, but that what any given student might learn is impossible to predict.

When the learning environment is well-regulated, much of the regulation is achieved 'upstream' of the lesson itself, through the setting up of didactical situations. The regulation can be unmediated within such didactical situations, when, for example, a teacher "does not intervene in person, but puts in place a 'metacognitive culture', mutual forms of teaching and the organisation of regulation of learning processes run by technologies or incorporated into classroom organisation and management" (Perrenoud, 1998 p100). For example, a teacher's decision to use realistic contexts in the mathematics classroom provides a source of upstream 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

---

[4]   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.

that this solution is unreasonable, and so the use of realistic settings provides a 'self-checking' mechanism.

On the other hand, the didactical situation may be set up so that the regulation is achieved through the mediation of the teacher, when the teacher, in planning the lesson, creates questions, prompts or activities that evoke responses from the students that the teacher can use to determine the progress of the learning, and if necessary, to make adjustments.

The 'upstream' planning therefore creates, downstream, 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 are constantly having to make sense of students' responses, interpreting them in terms of learning needs, and making appropriate responses. But they also arise when the teacher circulates around the classroom, looking at individual students' work, observing the extent to which the students are 'on track'. In most teaching of mathematics and science, the regulation of learning will be relatively tight, so that the teacher will attempt to 'bring into line' all learners who are not heading towards the particular goal sought by the teacher—in these subjects, the *telos* of learning is generally both highly specific and general to all the students in a class. In contrast, in much teaching in language arts and social studies, the regulation will be much looser. Rather than a single goal, there is likely to be a broad *horizon* of appropriate goals, 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.

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.

Second, 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 'upstream' regulation).

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 for the teacher herself, in that appropriate analysis of the students responses might suggest how the lesson could be improved.

Assessments can also be formative at the level of the school, district, and state provided the assessments help to regulate learning. Frequent assessment can identify students who are not making as much progress as expected (whether this is based on some notion of 'ability', prior achievement, or external demands made by the state). But frequent summative testing—we might call this micro-summative—is not formative unless the information that the tests yield is used in some way to change instruction (see next section).

System responsiveness and time-frames

A key issue in the design of assessment systems, if they are to function formatively as well as summatively, is the extent to which the system can respond in a timely manner to the information made available. Feedback loops need to be designed taking account of the responsiveness of the system to the actions that can be used to improve its performance. The less responsive the system, the longer the feedback loops need to be for the system to be able to react appropriately.

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

For an individual teacher, the feedback loops can be considerably shorter. A teacher might look through the same students' responses to a 'trial run' of a state test and re-plan the topics that she is going to teach in the time remaining until the test, such a test would be useful only weeks 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 just telling the teacher which topics need to be re-taught, but also to suggest what kinds of re-teaching might produce better results.

The building-in of time for responses is a central feature of much elementary and middle school teaching in Japan. A teaching unit is typically allocated 14 lessons, but the content usually occupies only 10 or 11 of the lessons, allowing time for a short test to be given, and for the teacher 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 key question given by the teacher 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 the 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.

However, this does not mean that the responsiveness of the system cannot be changed. Through appropriate 'upstream' regulation, the responsiveness can be enhanced considerably. Where teachers have collaborated to anticipate the responses that students might make to a question, and what misconceptions would lead to particular incorrect responses (e.g. through lesson study), the teachers would be able to adapt their instruction much more quickly, even to the extent of having alternative instructional episodes ready. In this way, feedback to the teacher that, in the normal course of things, might need at least a day to be used to modify instruction might be able to affect instruction immediately.

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

What does it mean for classroom assessments to be reliable? Valid?

As noted above, classroom assessment can contribute to both formative and summative assessment, although the bulk of this paper has dealt with the formative purpose.

For the summative purpose, it is clearly important that the scores, grade or whatever attached to students' work does not depend on the person doing the assessment in the sense that the assessment must be free from bias. This does not, of course, mean that all observers would necessarily agree on the score or grade, because the teacher will often have information from classroom observations or other sources to which other observers will not have access. In this sense the teacher is a privileged observer, not in terms of the standards they bring to bear, but in the access they have to information. A key requirement for classroom assessments to be reliable and valid, therefore, is that all the teachers involved in the assessment, across the school, district, state or even nation, have a shared construct of quality in respect of the particular curriculum philosophy in force for the subject in question. In other words, the teachers involved in the assessment must form a community of practice.

When assessment is conceived of in terms of the role it can play in the regulation of learning, the reliability and validity of an assessment relate to the learning that the assessment is intended to guide. Assessments would still therefore be reliable to the extent that they yielded consistent outcomes, but validity would relate to the inferences that the teacher drew about the learning needs of the student, and the necessary actions to take. In order to achieve this, each teacher must have not only a shared construct of quality, as required for summative assessment, but also an *anatomy* of quality. For a particular student response, whether written, spoken, or in even physical form (say a pitcher's pitching action), the teacher must not only be able to evaluate it as a whole, but also to be able to break it down into components to guide improvement.

However, a focus on the role of assessment in the regulation of learning also throws the role of the learner into clearer focus. For effective learning to take place, two further things must happen. First, the assessments must help the learners understand the standards of quality that the teachers already share. In other words, the assessments must help the students become members of the same community of practice of which the teachers are already members. Second, the assessments must help the students assess their own performance, and the performance of their peers, towards those goals.

Ultimately, the summative function is validated by the extent to which the meanings of assessment outcomes are shared. For the formative function, consistency of interpretation across teachers is less important than the consequences in terms of learning.

Conclusion

In this paper, I have argued that the terms formative and summative apply not to assessments themselves, but to the functions they serve, and as a result, the same assessment can be both formative and summative. Assessment is formative when the

information arising from the assessment is fed back within the system and is actually used to improve the performance of the system. Assessment is formative for individuals when they can use the feedback from the assessment to improve their learning. Assessment is formative for teachers when the outcomes from the assessment, appropriately interpreted, help them improve their teaching, either on specific topics, or generally. Assessments are formative for schools and districts if the information generated can be interpreted in such a way as to improve the quality of learning within the schools and districts. The view of assessment presented here involves a shift from quality control in learning to quality assurance. Rather than teaching students, and then, at the end of the teaching, finding out what has been learnt, it seems obvious that what we should do is to assess the progress of learning whilst it is happening, so that we can adjust the teaching if things are not working. In order to achieve this, the length of the cycle from evidence to action must be designed taking into account the responsiveness of the system. Some feedback loops, such as those in the classroom, will be only fractions of a second long, while others, such as those involving districts or state systems will last months, or even years.

More generally, I have suggested, building on the work of Philippe Perrenoud, that formative assessment be considered as a key component of well-regulated learning environments. From this perspective, the task of the teacher is to not necessarily to teach, but rather to engineer situations in which students learn effectively. One way to do this is to design the environment so that the regulation is embedded within features of the environment. Alternatively, when the regulation is undertaken through the teacher's mediation, it is necessary to build opportunities for such mediation into the instructional sequence by designing in episodes that will elicit students' thinking (upstream regulation) and to use the evidence from these probes to modify the instruction (downstream regulation).

References

Black, P.; Harrison, C.; Lee, C.; Marshall, B. & Wiliam, D. (2002). *Working inside the black box: assessment for learning in the classroom*. London, UK: King's College London Department of Education and Professional Studies.

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

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

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

Fontana, D. & Fernandes, M. (1994). Improvements in mathematics performance as a consequence of self-assessment in Portugese primary school pupils. *British Journal of Educational Psychology,* **64**(4), 407-417.

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.

Lesh, R.; Hoover, M.; Hole, B.; Kelly, A. E. & Post, T. (2003). Principles for developing thought revealing activities for students and teachers. In R. Lesh & H. M. Doerr (Eds.), *Beyond Constructivism: Models and Modeling Perspectives on Mathematics Problem Solving, Learning, and Teaching* Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

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

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

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

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

Wiliam, D. (1999, May) *A template for computer-aided diagnostic analyses of test outcome data.* Paper presented at 25th annual conference of the International Association for Educational Assessment held at Bled, Slovenia. London, UK: King's College London School of Education.

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. (2001). An overview of the relationship between assessment and the curriculum. In D. Scott (Ed.) *Curriculum and assessment* (pp. 165-181). Greenwich, CT: JAI Press.

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