Classroom assessment is not (necessarily) formative assessment (and vice-versa)

Paul Black and Dylan Wiliam

The terms 'classroom assessment' and 'formative assessment' are often used synonymously, but as the chapters in this collection show, the fact that an assessment happens in the classroom, as opposed to elsewhere, says very little, either about the nature of the assessment, or about the functions that it can serve. Classroom assessments may provide a sound basis for summative assessments, and those conducted outside the classroom may provide valuable insights into how to take learning forward. As well as the locus of the assessment, we think that it is also important to attend to the issues of authority, resources, interactivity and scoring. Each of these is discussed in turn below.

Authority. The assessments may be generated by the teacher, or by outside agencies, or, somewhere between these two extremes. In many European countries, assessments are proposed by the teacher, and approved (or not!) by an external agency, such as a regional inspector. Whether it is fair to assess students in different regions, or even students in different schools in the same region, on a different basis is, of course, problematic (see below).

Resources. The conditions under which students respond can be more or less controlled. At one extreme, typified by the traditional written examination, students may be required to respond alone, and without any additional materials. In other assessments, they may be able to consult specified textual resources (as in an 'open-book' examination), or a wider range of materials (for example, the internet) and even, in group projects, other students.

Interactivity. In the traditional test or examination, there is a stimulus, to which the student makes a response, which is then judged. There is no scope for the student to ask for clarification of the meaning of the stimulus, and the rater is required to make a judgement of the response as it stands. In this context, it should be noted that the majority of classroom tests or 'quizzes' that teachers employ in their classrooms, as part of their normal classroom practice, are of this sort. In an oral examination, however, the student can seek clarification of the meaning of the stimulus, and the rater can ask the student to clarify or elucidate their response. Furthermore, the oral examination allows the exploration of issues in depth, although this necessarily compromises the extent to which all students are examined on the same basis. In some examination systems, the ability to interact with candidates is regarded as essential for valid assessment. Of course, while face-to-face oral examinations have been the traditional way of providing for interaction between student and rater, modern technologies allow a much greater range of options, including the possibility of having computers, rather than humans, conducting the assessments.

Scoring: Where the results from the assessments are expected to serve summative or evaluative purposes, it is essential that the grades, marks or scores awarded depend as little as possible on who is doing the assessment—in other words that the assessments are *objective*. This is generally achieved by the use of machines, or by employing human scorers who have no knowledge of the student. At the other extreme, where assessments are intended to serve only a formative function, consistency of meanings across different raters

is less important. What is more important is whether the assessments lead to improved instruction. However, just as teachers can author assessments for summative purposes, they can also be involved in the scoring of their own students' work for summative purposes. One way that the necessary consistency of scoring across teachers has been achieved in the past is through scrutiny of the judgements made by assessors, which amounts to a kind of quality control process. Marks, grades or scores are generated, and at the end of the process, the quality of the assessing is inspected, and, if necessary, adjusted (this process is frequently termed 'moderation'). What is important here is that while the assessment may be conducted by the teacher, it is done so in a way that is inter-subjective—relying on the shared understanding of a community of teachers—so that the judgements are objective, in the sense they are free from individual subjectivity. Furthermore, the assessment is against a set of standards that are determined by the community rather than the teacher, so that even though the teacher is involved in scoring the student's work, the teacher is, in a very real sense, the student's ally rather than their enemy (e.g. "I'd love to give you an A for this but you just haven't reached the required standard yet").

A key theme running through several of the issues raised above is the extent to which all students are assessed on the same basis. Traditional wisdom dictates that fair assessment can be attained only if all students are assessed on the same basis, and this is the notion of 'fairness' used in traditional tests. However, in the same breath, it is also routinely acknowledged that it is essential to make adjustments to assessments for particular populations, such as students with visual impairments, specific learning disabilities (such as dyslexia) or motor impairments. At the higher levels of the educational system, it is routinely accepted that at least part the purpose of the assessment is to provide candidates with an opportunity to show what they can do, through the use of non-uniform assessments such as coursework, projects and theses. Of course, it could be argued that the requirement for non-uniform assessment at these higher levels arises from the complexity of the judgements that are necessary, but then the same also applies to earlier stages of the learning process—recent research has shown convincingly that the state of anyone's learning is a complex schema which defies simplistic analysis. Failure to recognize this (or, perhaps even worse, recognizing it but failing to acknowledge its importance) has resulted in a simplistic approach to assessment that leads to an emphasis on low-level aims that weakens validity. If more complex notions of fairness than simply making sure that all students are asked the same question are felt to be necessary for certain populations, then why not for all?

The chapter by Smithson and Porter provides a useful framework for beginning the process of examining the extent to which the (often noble) aims of standards are mirrored in classrooms, and in what is assessed. Given the prevalence of high-stakes assessment, we must accept that assessment may drive instruction, and therefore, there can be little hope of aligning instruction with standards unless the assessment is also aligned with the standards. The tools provided by Smithson and Porter can be used to illustrate, in a very convincing way, the extent to which these three aspects are aligned, and will frequently point to the need to improve the validity of the assessment being used.

In any well-designed system, the judgements of teachers will inform the summative function and external ideas about what to assess will inform instruction. This is particularly notable in the chapters by Forster and Masters (FM), Frederisken and White (FW), and Wilson and Draney (WD). What is at issue is the quality of the instruments and of the inferences made from them.

The approaches outlined by FW and FM signal a move away from a traditional quality control orientation towards one of quality *assurance*. The major effort goes not into correcting marks that are wrong, but into improving the ability of the assessors to get it right first time. It is also noteworthy that both FW and FM focus on securing consensus not through getting teachers to agree on some lowest common denominator, but through beginning to address explicitly the features that are likely to be present in good responses. The notion of 'community of practice' is a useful idea for thinking about how teachers can come to consensus over the marks, grades or scores to be awarded to students' work, but can serve to disguise what it is that they come to agree *on*. After all, the requirements of reliability are met if teachers' judgements are consistent, even if they have no idea what they are doing, or how they are doing it. The result of this can often be that teachers can judge accurately the standard of students' work, but have little idea about how to improve it.

Our own work with teachers and students suggests that when teachers take pains to share with their students the criteria for success, that students are able to internalize these quickly (and often more quickly than the teachers would have thought possible). As Royce Sadler (1989) notes, this is a necessary, but not sufficient, condition for improvement. For progress to be possible, in addition to having a notion of quality, it is necessary for either the teacher or the learner to have an *anatomy* of quality. In other words, it must be possible to break down the path from the current position to the goal into a series of steps that the learner can take. While this may be possible for the learner to do for themselves, more often it will be the responsibility of the teacher, and our experience is that many teachers do not have very clear models for progression. As one teacher on an in-service program remarked, in the context of the National Curriculum for English, "I know he's a level 4, but I don't know what to do to get him to level 5." This is why the developmental models inherent in the assessments described in FW, FM and WD are so important. As well as grounding assessment in developmental principles, they provide models for teachers to help them understand the nature of progression in the domain, and thus support them in identifying 'next steps' for students.

In discussions of assessment, it is commonplace to distinguish between assessment *of* learning and assessment *for* learning. However, the rich tasks proposed by FM, FW and WD demonstrate an intermediate possibility—assessment *as* learning. While the tasks proposed in these three chapters have as their primary goal the rendering of accurate, meaningful judgements of students' achievements, the tasks themselves are valuable learning activities. The extended nature of these tasks enhances reliability, but the extra time that these tasks take is justifiable only because students are learning while they undertake them. These activities therefore assess not what the students know at the beginning of the assessment, but at the end.

For assessment to function formatively, we need accurate information about where students really are in their learning, but this is also what we want for summative assessments. It may be therefore, that we can find synergy rather than tension in the relationship between formative and summative. For example, consider the following item:

Give an algebraic expression for the nth term in the sequence 5, 8, 11, 14...

A) $n + 3$

B) $5 + n$

C) $3n + 2$

D) $2n + 3$

This item is designed to support both summative and formative inferences. The distractors here are derived from well-known difficulties that students encounter in this domain. Distractor (A) is attractive to students who have established the term-to-term rule (i.e. add 3) rather than the position-to-term rule that is required. Distractor (B) is also based on a recursive approach, but focuses on the idea of adding a number to the initial term.The key (C) and the last distractor (D) permute the parameters.

Thus in putting this question before a student the teacher will have in mind a tentative judgement of the pupil's understanding, and will be able to confirm or amend this judgement in the light of the response. In addition, choices such as (A) or (B) will serve to pin-point different types of error. A student choosing (A) may have misunderstood the requirement of the question for a general term rather than just a rule for generating the sequence. The student who chooses (B) is likely to believe (as they have been repeatedly told!) that "n can be any number", and this misconception will need to be addressed (note that for this distractor, it is important that it is expressed as $5 + n$ rather than the more usual $n + 5$). This item could therefore serve both formative and summative functions well. The diagnostic potential of this item could be improved further by changing distractor (D) for an alternative offering "n3 + 2". Then, where formative functions are paramount, all of the distractors will be useful. However, many would argue that the "n3 + 2" distractor is unfair in the context of a high-stakes assessment—especially if it were to be scored as incorrect.

All of the papers embrace, explicitly or implicitly, the idea that we must find ways of integrating the formative and summative functions of assessment, but more work needs to be done on understanding the nature of the relationship between the two. If such an integration is to be found, it will certainly not rest on a simple equation of external assessment for summative purposes on the one hand, and classroom assessment for formative purposes on the other.

Reference

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