Part I

Lighting the fuse

1 Assessment to support learning

Wynne Harlen and John Gardner

This chapter provides an overview of the purposes and uses of classroom assessment by teachers or teacher assessment as it is better known (not to be confused with the assessment of teachers). At the heart of the matter lies the distinction between formative and summative uses of assessment information, and the basic argument that teacher assessment may be successfully used in both contexts. A primary aim of this chapter is therefore to clarify the often confusing terminology that misleadingly promotes the view that the method of assessment has to be exclusively formative or summative. School, that is classroom-based teacher assessment is compared to external testing in terms of the different roles they play and the advantages and disadvantages they each have.

In 2007, Newton argued that assessment processes of different kinds may serve many purposes, in at least 22 categories that he illustrated in his paper. They included the familiar formative, diagnostic and qualifications-related uses along with less familiar variants such as placement, licensing and programme evaluation. In common with others in the field, he rejects the simplistic notion of an assessment being itself formative or summative, arguing that it is the purpose to which an assessment is put that is the crucial distinction. To try to limit inappropriate use of assessment information, he argues that anyone developing any form of assessment should identify the purposes for which it is fit and most importantly the purposes for which it is unfit: 'Stakeholders should be deprived of ignorance as an excuse for misuse' (Newton, 2007: 168).

There are strong demands for precise use of assessment terms in Newton's work and the nuances of meaning are well worth pursuing. However, we are interested here in the most practical of purposes, the use of assessment to support improved learning. Assessment by teachers in the classroom; that is, teacher assessment, has many purposes to which it can be put but our focus is on the most obvious: for helping pupil to learn and for contributing to judgements on pupils progress and achievement.

Purposes

The title of this chapter, *Assessment to support learning*, could well be taken to mean that it is about formative assessment, or assessment for learning. This purpose of assessment is now commonly accepted, if less widely practised, than it is often assumed to be. It is sometimes contrasted with summative assessment, or assessment of learning, which serves the purpose of reporting on what has been learned at a particular time. These are the two main purposes of assessment and there are conflicting claims about how distinct they are in practice. For instance, in the view of Gipps (1994), 'Any attempt to use formative assessment for summative purposes will impair its formative role' (p. 14), while Black et al. (2003) advocate the use of summative tests to help learning. Certainly the study of current assessment practice leaves no doubt that summative assessment has an impact on learning but whether this is positive or negative depends on the use made of the information gathered in the assessment.

Uses

In the case of formative assessment there is one main use for the information – to help learning. Assessment that does not do this simply cannot be called 'formative'. By contrast, information from summative assessments can have various uses. The two main uses fall under the headings: use of results for individual students and use of results for groups of students (classes, year groups, national populations). Within the school, individual students' assessment results may be used for record keeping, monitoring progress, reporting to parents, students and other teachers, and for career guidance. Summative assessment of individual students may also be used by agencies outside the school to select students or to award qualifications. Both of these uses directly affect the individual student to some degree. In addition, the aggregated results of summative assessments for groups of students are used both within and outside the school. Within the school they may be used for school self-evaluation, to monitor trends in performance or perhaps to evaluate the impact of changes in procedures or resource materials. Perhaps more controversially, aggregated results may also be used by agencies and authorities outside the school for:

- accountability – evaluation of teachers, schools, local authorities against targets;
- monitoring – students' average achievements within and across schools in particular areas, and across a whole system for year-on-year comparison.

Both of these external uses are problematic, particularly when the only information used is derived from test scores. Although the main focus of this book is assessment for uses relating to individual students, this use has to be seen against the background of the same data being used for accountability and particularly for creating performance targets and league tables of schools, as has been the practice in England since the introduction of national testing. As reported by Harlen and Deakin Crick (2003), this puts teachers under pressure to increase scores, which is widely recognized as leading to teaching to the tests, giving multiple practice tests and coaching pupils in how to answer test questions rather than in using and applying their understanding more widely. Other known consequences have been charted by the Assessment Reform Group (ARG, 2002a) as the demotivation of lower-achieving pupils and, for all pupils, a view of learning as product rather than process. It also leads to undue attention being focused on those who are performing just below the target level, with less attention for those who are either too far below or are already above the target level.

These effects are by now widely known and recognized by students themselves: 'Students are drilled to jump through hoops that the examiner is holding . . . The mechanical exam process is moulding a mechanical education' (Tom Greene, a secondary school pupil, writing in *The Independent*, 17 August 2006). – and by parents:

> For my son, and for most 10-year-olds in the country, the next nine months will be . . . a sterile, narrow and meaningless exercise in drilling and cramming. It's nothing to do with the skills of his teacher, who seems outstanding. Nor do I blame the school. It's called preparing for Key Stage 2 SATs.
>
> (Benaby, 2006)

as well as by teachers (NUT, 2006) and researchers.

For monitoring standards of pupil achievement at the regional or national levels, the interest is not in the performance of individual students but in the population performance in each learning domain, such as different aspects of mathematics, reading or other subjects. Thus, validity depends on how well the domain is sampled. If the data used in monitoring are derived from a summation of individual test results, as is the case in England where national tests results are used to monitor change in national standards, then the sample of the domain is restricted to the questions that any individual pupil can answer in a test of reasonable length. The questions do not necessarily represent a good sample of the domain, and will depend on the particular content of the test. Monitoring in this way does not provide sound information about changes in national levels

of achievement, yet these are taken as measures of national 'standards' and important policy decisions are based on them.

We discuss the meaning of 'standards' in Chapter 2, but note here that a more valid approach to monitoring achievement levels would be to use a far greater number of items, providing a more representative sample of the domain. Since the concern is not with the performance of individual pupils, there is no need for all pupils to be given the same items. All that is needed is for each item to be attempted by an adequate sample of the population. Sampling of this kind, where only a small proportion of students is selected and each only takes a sample of the full range of items, is used in international surveys. These include the Programme for International Student Achievement (PISA) (OECD, 2009) and the Trends in International Mathematics and Science Study (TIMSS) (IEA, 2009) while national surveys include the Scottish Survey of Assessment (SSA) (SEED, 2005) and the older Assessment of Performance Unit (APU) for England, Wales and Northern Ireland (Foxman et al., 1991). As outlined in Chapter 3, there are emerging signs that the APU-like sampling strategy is being reconsidered for operation in England and Northern Ireland as a means of establishing national standards without burdening pupils unduly.

Impact on learning

As mentioned in the introductory chapter and above, formative and summative assessment are not different kinds of assessment but do serve different purposes. Indeed, there is an argument that the terms 'formative assessment' and 'summative assessment' should not be used and should be replaced by 'assessment for formative purposes' and 'assessment for summative purposes". However, so long as it is understood that the information from any assessment process has the potential to be used directly to help learning or simply to judge or record it, the former terms are useful in common parlance.

A teacher's mark or grade on a student's work may therefore end up simply as another entry in a record sheet or it may be used as a start of feedback to the student. For example, it could help with a pupil's future work by explaining the criteria used in arriving at the mark and identifying what aspects of the work were taken into account in the judgement. Hopefully, even the mark in the record book will also be used in a future review of the student's progress with a less direct, but still positive, role in providing for future learning. A student's participation in a national or international survey appears to have no possibility of improving that student's learning, but in the longer term the results of the survey should be used to improve the learning opportunities of other students, since

the data collected make it possible to relate test results to the conditions of learning. If there is no prospect of the results being used to support learning, it is hard to understand why governments provide the quite considerable funding that such surveys require.

These arguments lead us to conclude that all assessment should ultimately improve learning, at the level of the individual student or through changes at the system level. This is one of the overarching principles that we develop in this book. It applies as much to assessment that is summative as to formative assessment, where it is a necessary requirement. The aim ought to be to conduct assessment for summative purposes in a way that supports the achievement of all learning goals and does not limit attention only to those learning outcomes and processes that are easy to assess. This is likely to mean that various ways of collecting assessment data are needed. Some outcomes, such as basic numeracy and literacy, may be quite adequately addressed by well-designed tests, but our concern here is with those that are not. When tests are favoured in government policies and are given high stakes by their use in setting targets and evaluating teachers and schools, the outcomes not adequately assessed by tests are in danger of being neglected in teaching.

The formative role of assessment

When assessment is specifically intended to help learning, a simple check on its effectiveness would be to assess whether it leads to greater learning than would be the case when it is not being used. However, this is not a realistic criterion to use as it begs a number of questions about the nature of the learning, the period of time it covers and the competition with other influences on learning and so on. Indeed, it is not necessarily a straightforward matter to say whether the formative use of assessment is actually under way, given its complexity. Most projects aiming to improve formative assessment in normal classroom conditions do not provide research evidence that enables these questions to be adequately addressed.[1] Instead, the case for the importance of formative assessment can be made from arguments based on what is known about learning and from the evidence derived from controlled research studies, which show that improved learning follows when certain features of using assessment formatively are in place.

Arguments based on learning

Current views of learning emphasize the importance of the active role of learners in developing their understanding through using existing ideas

and skills to build 'bigger' ideas and more advanced skills. 'Big' ideas are ones that can be applied in different contexts; they enable learners to understand a wide range of phenomena by identifying the essential links ('meaningful patterns' as Bransford et al., 1999, put it) between different situations without being diverted by superficial features. Merely memorizing facts or a fixed set of procedures does not support this ability to apply learning to contexts beyond the ones in which it was learned. Knowledge that is understood is therefore useful knowledge that can be used in problem-solving and decision-making.

The teacher's role in this learning is to make provision for appropriate experiences that challenge students' existing ideas and skills but are not out of their reach. It means that teachers take steps to find out what sense students are making of their learning activities. Particular kinds of 'deep' questioning are important here and students need to have opportunities to reflect, discuss and consider thoughtful answers to such questions. The information gathered by teachers has to be fed back to students and used to regulate teaching so that the pace of moving towards learning goals is adjusted to ensure the engagement and active participation of the students. Students can participate in these processes if teachers communicate to them their goals for the lesson and the criteria by which they can judge their progress towards the goals. The lesson goals may well vary in detail for different students according to their previous experience and progress in their learning.

It follows, then, that some of the key features of the formative use of assessment are likely to be that:

- information about ongoing learning is used in decisions about the pace and content of teaching;
- teachers ask questions that enable them to know about the developing ideas, skills and attitudes of their students;
- students are provided with feedback in a form that helps them engage with further learning;
- students take an active part in their learning and participate in deciding the goals to which they are working;
- students understand the quality criteria to be applied to their work and so can take part in self- and peer-assessment.

Research evidence

These points emerge from considering how learning with understanding takes place and reflect closely the features of classroom assessment that research has found to improve learning. Indeed, there is a large and growing body of evidence that formative assessment improves learning. Empirical investigations of classroom assessment have been the subject of several

reviews of research, principally those by Natriello (1987), Kulik and Kulik (1987), Crooks (1988), Black (1993) and Black and Wiliam (1998a). The last of these has attracted a good deal of attention world-wide for several reasons. For example, its dissemination in the form of a short booklet, *Inside the Black Box* (Black and Wiliam, 1998b), meant it has reached a much greater audience than would a research journal article. In addition, the authors' quantification of the positive impact on learning, of using assessment for learning (AfL), made compelling reading. They estimated that the gains in learning were large enough to 'improve performances of pupils in GCSE by between one and two grades' (Black and Wiliam, 1998b: 4). Further, they reported that 'improved formative assessment helps the (so-called) low attainers more than the rest, and so reduces the spread of attainment whilst also raising it overall' (ibid). This means of expressing the impact of AfL was more effective in communicating with policy makers than providing evidence in the conventional academic manner.

Black et al. (2003) cite research by Bergan et al. (1991), White and Frederiksen (1998) and the review of Fuchs and Fuchs (1986) as providing evidence of better learning when teachers take care to review information about students and to use it to guide their teaching. Butler (1988), for example, showed the importance of non-judgemental feedback in the form of comments with no marks. Schunk (1996) also found positive impacts on achievement as a result of students' self-assessment. These all reflect a view of learning in which students participate actively rather than being passive receivers of knowledge. This means that assessment, which is used to help learning, plays a particularly important part in the achievement of the kinds of goal of understanding and thinking valued in education for the twenty-first century.

Extent and quality of practice

The practice of using assessment to help learning is a complex combination of interconnected features as listed above. However, no study to date has attempted to combine all the aspects of formative assessment that have been the subject of separate studies. It is therefore not clear just how important it is for all features to be in place for the benefits to be realized. Taking on change in the full range of practices is likely to be overwhelming and most teachers begin by choosing one or two changes, such as in questioning or procedures for feedback to students. More information is needed to confirm whether this selective approach can be regarded as effective formative use of assessment.

What is already clear, however, is that positive impacts may not result if teachers are following procedures mechanically without understanding

their purpose. For example, Smith and Gorard (2005) reported that when some teachers put comments but no marks on students' work (a key feature of the formative use of assessment), these students made less progress than others given marks. Most notably, however, the majority of the teachers' comments illustrated by Smith and Gorard merely commented on how well the students had done and did not supply guidance on how to improve the work. Almost a year later, after a systematically more thorough programme of introducing AfL, including support for written feedback, the assistant head of the same school was reporting that Key Stage 3 results for English, mathematics and science had steadily improved (Burns, 2006). It is essential, therefore, for teachers to understand the underlying rationale for any new approach, and to embrace the change in teacher – student relationships that is involved in using assessment to help learning. Without this change, students will not use feedback to improve their work or to reveal their understanding and difficulties, as is necessary for assessment to be used for learning. This example underlines the point that it is necessary to distinguish between bringing about change in assessment practice and bringing about change that is consistent with improving engagement in learning. It also emphasizes that such change will not happen without specific support to help the teachers to assimilate the underlying rationale for the change.

There is, therefore, a matter of quality of practice to be considered, which involves a general change in the relationship between teacher and students. Wiliam et al. (2004) quote Brousseau (1984) who describes this as renegotiating the 'learning contract' between teachers and students. This refers to the shift in responsibility for learning from belonging only to the teacher to being shared with students. As Harlen (2006) has argued, openness in relation to assessment also provides the context for assessment evidence, gathered and used as part of teaching, to be the basis for the summative role of assessment of learning outcomes relating to all learning goals.

The summative role of assessment

The impact of assessment

There is a good deal of evidence to support the claim that outcomes that are assessed are the ones that are given most attention in teaching, particularly when high stakes are attached to the results. Harlen and Deakin Crick (2003) have shown that setting targets in terms of test results, with sanctions attached to failure to meet the targets, leads to a range of practices that narrow students' learning experiences and affect their motivation

for learning. In James et al.'s (2006a) study, evidence from 1,500 staff in 40 primary and secondary schools in England led to the conclusion that there is no doubt that teachers are teaching to the tests their pupils have to take; they do not feel they can do anything else. Marshall and Drummond's (2006) case studies confirmed this view, revealing that teachers believed that 'there are circumstances beyond their control which inhibit their ability to teach in a way they understand to be good practice' (p. 147). These 'circumstances' are the tests, which cannot, on account of their form and length, adequately assess certain important learning outcomes. Learning experiences that lead to deep understanding are unlikely to receive the attention that matches the rhetoric of such learning unless it is included in the assessment that matters.

The limitations of tests

The primary reason that these learning outcomes are not currently included in the assessment of learning outcomes that 'matter' relates to the difficulty of assessing them through the methods that are presently favoured. For assessment where the results are used beyond the school, tests are preferred because they are considered to be of high reliability and to be 'fair'. In this sense, reliability refers to whether the result would have come out differently on a different occasion for the same student assessed on the same learning outcomes. However, as the work of Wiliam (2001), Black and Wiliam (2006), and Gardner and Cowan (2005) has demonstrated, even if every item were to be free of ambiguity and could be marked without error, there is still an unavoidable error in the overall test score. For example, error is introduced when the items included are a selection of possible choices and a different selection could produce a different result. In practice, items can never be error-free and the steps taken to raise reliability favour items that are 'closed', that is, having fixed responses. Their marking then depends as little as possible on human judgement. Clearly, items that require students to be creative, present arguments or show understanding of a complex situation do not fit this description and appear less frequently in current tests than they ought to. Even when such items are included, high stakes add pressure that leads to teaching how to pass tests rather than spending time helping students to understand what is being tested. As Harlen and Deakin Crick have argued:

> Direct teaching on how to pass the tests can be very effective, so much so that Gordon and Reese (1997) concluded that students can pass tests 'even though the students have never learned the concepts on which they are being tested'. As teachers become more adept at this process, they can even teach students to answer

> correctly test items intended to measure students' ability to apply, or synthesize, even though the students have not developed application, analysis or synthesis skills. Not only is the scope and depth of learning seriously undermined, but this also affects the validity of the tests, for they no longer indicate that the students have the knowledge and skill needed to answer the questions correctly.
>
> Harlen and Deakin Crick (2003: 199)

As implied here, validity refers to how well the assessment reflects the achievements that it is intended to assess. If students are under so much pressure that they feel anxiety due to the high stakes associated with their performance, they may not perform as well as they are able. This may act to reduce the validity of the test. The extent to which an assessment is capable of capturing the achievements intended can be established by expert judgement – for instance in comparing the content of the tasks assessed with the content of the curriculum or course it is designed to assess – or by statistical examination of consistency in the results. However, the concept of validity is rather more complex than this, particularly because it will depend on the use made of the results. For instance, an examination result may be reasonably valid for assessing achievement as a result of a course, but less so for predicting achievement in further courses. In order to decide the most valid data for a particular use, it is necessary to consider the needs and points of view of those who use the information.

The potential of tests

It is worth noting that valid items assessing such outcomes as problem-solving, enquiry skills and critical thinking can be created but, because the response to them is highly dependent on the choice of content, a large number of items is needed to obtain a reliable score. Thus, they can be used in surveys where every individual does not need to take every item. Some items of these types were included in surveys conducted nationally by the APU in England, Wales and Northern Ireland in the 1980s. They also currently feature in the National Assessment of Educational Progress (NAEP) in the United States and the (SEED, 2005). The items in these surveys are not limited to what can be assessed on paper and include, for example, assessment of listening and speaking and performance of investigations in science and mathematics. As they are designed to monitor performance of a population in a particular learning domain, the results have no significance for individual students or even classes and schools. They are, therefore, low stakes and there is no incentive to teach to what is

being tested. The influence of the test content is at the conceptual level of the test framework, not in relation to specific items. As Kellaghan (1996) argues, most learning from the surveys is at the system level where the results inform policy makers not only about performance and trends in performance across a range of aspects within each domain tested. They also provide information about how factors such as curricula, time spent on school work, teacher training and class size are found to be associated with variation in student achievement.

Possibilities offered by teacher assessment

Returning to the matter of assessing individual students, it is evident that tests of any reasonable length are not reliable or valid for assessing certain learning outcomes. In particular, they are not suited to assessing some of the essential elements of twenty-first-century education such as problem-solving, critical thinking, enterprise and citizenship. Valid assessment would require students to be in situations where they can demonstrate these attributes when they are assessed; faced with real problems and required to link one experience with another. An alternative to written tests is clearly needed if we are to include these outcomes in summative assessment. It can be found in using the judgements of the teacher, acknowledging that the experience students need in order to develop the desired skills, understanding and attitudes also provide opportunities for progress towards these outcomes to be assessed. Assessment by teachers can take evidence from regular activities, supplemented if necessary by evidence from specially devised tasks; that is, introduced specifically to provide opportunities for students to use the skills and understanding to be assessed.

Over the period of time for which achievement is being reported (a term or half year for regular reports to parents and one or more years for external certification), students have opportunities to engage in a number of activities in which a range of attributes can be developed. These same activities also provide opportunities for the development to be assessed by the teacher. In other words, the limitation of the restricted time that a test provides does not apply when assessment is teacher-based.

Teachers' assessments are often perceived as having low reliability but the evidence for this comes from situations and studies where no moderation or other form of quality assurance has been in place. Clearly, some quality assurance of the process of arriving at judgements is necessary, particularly when the results are used for decisions that affect students' future learning opportunities. According to Harlen (2004), when steps are taken to moderate the results, the reliability of teachers' judgements is

comparable with that of tests. Moreover, the moderation process is itself widely recognized as being a valuable form of professional learning. For example, Maxwell, referring to experience in Queensland, comments that:

> The most powerful means for developing professional competence in assessment is the establishment of regular professional conversations among teachers about student performance (moderation conversations). This is best focussed on actual examples of student portfolios.
>
> (Maxwell, 2004: 7)

Advantages of teacher assessment

There are three further advantages of using teachers' judgements for the summative use of assessment. The first is that it enables processes of learning as well as outcomes to be assessed. Such information is particularly useful where the ability to undertake further learning is of interest. For example, for those who select students for advanced vocational or academic courses of study, it is as important to know if candidates have developed the skills and desire for learning, that is, if they have 'learned how to learn' and are likely to benefit from further study.

The second advantage is that using teachers' judgements opens the possibility of students playing some part in the assessment of their learning outcomes for summative purposes. This requires that they know the criteria by which their work is judged, taking away the mystery and anxiety often associated with some assessment procedures. The criteria ought to be progressive so that students see not only what they have achieved but what they have still to achieve. Students need also to be made aware of the purpose of the assessment and how it can help them to recognize their strengths and where they need to make more effort. This enables the process of arriving at a summative judgement to be used formatively by students, who see what they have to aim for in the longer term, and by teachers as feedback into planning.

The third advantage is that evidence collected and used formatively can be used for summative purposes when judged against the standards for reporting students' levels of achievement. The mechanism for doing this, however, must take account of the differences in the nature of the judgements made for formative and for summative purposes. Evidence collected and used formatively is detailed, relates to specific lesson goals and will be used to give positive feedback. It takes into account the particular circumstances of individual students and assists in making judgements

about next steps in learning. It leads to action and not grades or levels. For summative use, the evidence from formative assessment needs to be brought together and judged against the criteria that indicate the various grades or levels used in reporting. Thus, the evidence can be used for two purposes, with the proviso for summative use that it is reinterpreted against the same reporting criteria for all students. This involves finding the 'best fit' between the evidence gathered about each student and the reporting levels. In this process the change over time can be taken into account so that preference is given to evidence that shows progress during the period covered by the summary judgement or report.

Conclusion

Externally generated and marked tests and tasks have important roles to play in schools, for example, in helping teachers to benchmark their understanding of levels of performance. However, their use for supporting learning is less obvious. More contentious uses such as school and teacher evaluation do render them problematic in a variety of disruptive ways such as their impact on what is taught, the targets they give rise to and the burdens of anxiety and time that they may place on the learning process.

In contrast, teacher assessment comprises a large collection of information gleaned from the daily classroom interactions between pupils and teachers, and between pupils and pupils. The interactions cover many different types of process including the dynamic assessments of questioning and feedback, the reflective assessments of self- and peer-assessment, the sharing of learning goals and the criteria that indicate their achievement, and the long-term progression-related evidence from pupils' work. Such a wealth of evidence is primarily used in an *ad hoc* support of learning 'in the moment' (assessments for formative purposes) but can also be captured in suitable forms for reporting progress and performance (assessment for summative judgements). As Brooks and Tough put it:

> The most effective schools now practise a culture of continuous teacher-led assessment, appraisal and adjustment of teaching practices to personalise learning for all their pupils. It seems clear that assessment that does not assist the learning of the child is of very limited value, and in many ways the certification of achievement and the accountability role of assessment are only important because of their links to this.
>
> (Brooks and Taylor: 2006)

Questions for reflection

1. In what contexts is it important or not important that classroom-based teacher assessment should be made as reliable and valid as possible? What are the reasons for this?
2. Why does our education system require summative judgements to be made on pupil progress and performance? Does this requirement compromise the use of assessment to support learning?
3. What might be the arguments for and against the use of surveys to provide information at a national level?
4. What might a system look like where all assessment supported learning?

Note

1. The work on outcomes of the King's Medway Oxfordshire Formative Assessment project (KMOFAP) reported by Wiliam et al. (2004) is an exception. See Chapter 8.