Chapter 7

Teachers' Roles in Assessment and Testing

Introduction

It is clear that in formative assessment, the central role belongs to teachers. It is arguable whether or not they also have a role in assessment for certification or for accountability. At the same time the following quotation, written in 1986, underlines an important feature:

> Consider the amount of time, energy and money spent by both individual teachers, and schools in general, on setting and marking continuous assessment tests, end of session examinations and mock 'O' levels. Reflect on the money spent by examination boards and the number of assessment specialists employed by them. Read, if you can find a sabbatical term, the literature on the technology of assessment for reporting and certification. Compare these in turn with the complete lack of support normally given to teachers in devising and applying procedures to pinpoint their students' learning problems, with the virtual absence of outside agencies to develop formative assessment instruments and procedures, and the limited literature on the topic. (Black, H. 1986, p. 7)

This neglect seems all the more surprising in the light of three significant considerations. These are:

- It seems self-evident that good quality feedback is essential to effective teaching and learning.
- There is well documented research evidence that formative assessment linked to good learning practice does indeed give improved learning (Black, P.J. 1993(a), pp. 54–6; Crooks, 1988).
- Assessment by teachers can make a strong contribution to overcoming some of the serious problems of obtaining valid assessments for certification and for accountability.

The presentation in this chapter will start with a discussion of formative assessment and its close link to the improvement of learning. This will establish some principles, in the light of which the next section can review the present state of practice, taking up some of the implications of the situation

described in the above quotation. This will lead naturally to a section on techniques for developing formative assessment. The discussion then turns to the role of teachers in summative and certification assessment, followed by a section on the tensions between the formative and summative roles. A final summary reviews implications for the future.

Evidence for Learning

Formative assessment requires the use of a diverse set of data for a purpose. That purpose is the modification of the learning work to adapt to the needs that are revealed by the evidence. Only when assessment evidence is acted upon in this way does it become formative. Such response to feedback may range from the immediate classroom response to a question, through to a comprehensive review of a variety of data in order to appraise progress over a whole topic or theme. In terms of control theory, the use of feedback in this way would seem to be an obvious necessity.

Assessment of a student's response to any activity involves an interpretation of that response. For example, the teacher concentrating on the formative purpose will be looking for evidence of any obstacle of misunderstanding with which the student needs help, whereas for the summative purpose, the task is to determine the extent to which the student's work has met given target criteria (Wiliam and Black, 1996).

Authentic Evidence

It follows from the findings about 'Validity and learning' discussed in Chapter 4 that the design and interpretation of assessment work has to be conducted within a strategy concerned with the development of the active and thoughtful learner. Such active learning is quite different from passive reception learning, and calls for a broader range of assessment activities. This point is illustrated by the following comment from a teacher in Spain:

> . . . this diversity of strategies for the assessment of the learning process owing to the different forms of communication that it provides for the pupils, offers the latter a greater chance of performing well in their studies. In the case of those students who have highly-developed instrumental and operative abilities or mathematical intuition, the exercises and problems based on physics and chemistry formulas enable them to demonstrate what they have learned very easily. Likewise, those students who display a more rational, divergent way of thinking, who interpret things in a more visual manner and who have powers of analysis and synthesis are able to show what they have learnt by means of . . . their activity books, reports on experiments, hypotheses based on problem solving, correlation of data and the interpretation of graphs and charts. (Black and Atkin, 1996, p. 112)

One of the tasks of the teacher is to find a language or medium in which the pupil can respond. This responsiveness is to be assured by having both diversity and flexibility in the assessment process (Darling-Hammond *et al.*, 1995, p. 258). As pointed out in the discussion of 'Authentic assessment' in Chapter 6, classroom assessment can use a far wider range of assessment evidence than is possible for external assessments, and such a range has to be used if a valid picture of a pupil's attainments and needs is to be obtained. Furthermore, in interpreting pupils' work for any of the purposes, the teacher can work in a manner different from an external examiner, as illustrated by the following extract from an account of a teacher in the Netherlands engaged in teaching a new curriculum in technology:

> The assessment of practical activities is more complicated. Most teachers also take the way in which a piece of work has been produced into account. In other words, they evaluate both the result and the way in which it was produced. The teacher's evaluation of work pieces is personal, though it has tried to use objective criteria (for example, a comparison of a drawing to guidelines in the assignment). Some teachers involve other pupils in the assessment. (Black and Atkin, 1996, p. 96)

The justification for taking into account the way in which a piece of work was produced is that any inferences drawn from the assessment result about the student's skill and understanding may be more trustworthy if this is done — thus enhancing the validity of the assessment (as defined in the quotation from Messick in Chapter 4).

The teacher's personal knowledge of the pupil, and understanding of the context of the performance, are significant advantages in securing validity in formative assessment. A further advantage arises because the closer an assessment activity can come to the actual activity to which its results are to be considered relevant, the more likely it is to satisfy validity criteria. As explained in Chapter 6, the term 'authentic assessment' serves to signify one of the strengths of classroom assessment. In striving for validity, such assessment has a better chance of success than formal timed written tests. Such formal tests are inevitably presenting a situation which is artificial and there is good evidence that such 'surrogate' forms of testing cannot serve as substitutes for authentic assessments.

Given that the purpose of formative assessment is to identify particular learning needs, it has to be criterion referenced. More precisely, it has to be domain referenced. However, the difficulties that beset domain referencing (as discussed in Chapter 4) are not so serious in the formative context, because on any one occasion only a small and well-defined element of a domain will be at issue in learning, and because there is opportunity for a diverse and flexible approach (Black and Dockrell, 1984, pp. 116–18).

The need for an approach in which activity is assessed in context is argued at a more fundamental level by Christie (1995, p. 109) in an analysis

(already referred to under 'Validity and learning' in Chapter 4) of the problems of assessing the domain of reading. He proposes three potential dimensions along which a pupil's progress should be followed: maturity, skills and competencies, and contexts of performance. Any standard test designed for reliability by way of internal homogeneity is bound to focus on only one dimension, which may make the evidence useless — as Christie explains:

> . . . attempts to manipulate a predictive variable, rather than to manipulate the performance that the variable predicts, immediately destroy the set of relationships which lent the variable predictive validity in the first place. (p. 112)

To explore and work at all three requires a more flexible approach in which each piece of work is judged in terms of its particular combination of qualities in these dimensions, and guidance for a pupil would have to look at assessing and challenging progress in all three dimensions. Such an emphasis leads to the conclusion that the data that will guide in a formative way will be qualitative descriptions of the particular combination of qualities in a piece of work rather than a single number or grade.

Resources for Diagnosis

Whilst teachers' own assessments have to be closely coordinated with their teaching plans, it does not follow that external agencies have no role in supporting such work. In some countries, national assessment programmes have been based on the provision of assessment items for teachers to use as and when they find helpful — in some cases both to use the results internally and for reporting to outside agencies, so that the assessments serve both formative and summative purposes (Harlen, 1995). External provision can suggest useful frameworks for formative assessment as well as supplying test exercises for formative use. Standardized tests are unlikely to contribute because they tend to be norm-based and lacking in the detail required for useful feedback. Sets of more open questions can suffer from the same drawback of being too general, even though they may do a good job for an examiner trying to cover a field comprehensively with as few questions as possible. Where questions have been composed to serve a formative function, as in light-sampling surveys, so that the questions can cover a field more thoroughly and in detail, the resource can be of considerable value.

A national innovation in France has taken this further by setting up a requirement for national tests to be taken by all students at three stages — 3rd Grade (age 8, middle of primary school), 6th Grade (age 11, beginning of lower secondary) and 10th Grade (age 15, beginning of upper secondary). The significant feature is that these tests are taken at the beginning of

a new school year, so that the emphasis is not on the past, but on diagnostic information at the outset of their school year to help the design of teaching for each class of pupils. A detailed evaluation of this initiative has been carried out for mathematics (Black and Atkin, 1996, pp. 103–8). This study revealed some important general lessons.

The innovation required the construction by experts of test items which were designed to explore critical features of students' understanding of mathematical concepts, i.e. they focused on subject competence and on cognitive development. Thus, national resources were being used to strengthen teachers' own work. This required however extensive training of the teachers, because they had to understand the strategies behind the questions in order to mark and interpret the responses in terms of critical learning needs of their students.

The information that such detailed assessment could reveal was clearly worthwhile. However, the evaluation of this innovation revealed some other important lessons about this approach to formative assessment. One was that the diagnostic information has to be up-dated frequently — as the evaluation report noted:

> While the results supply quick and useful information on a class, this information is perishable and rapidly outdated. The teachers point to not only the fleeting character of the information obtained from the results (the term 'snapshot' was more than once used) but also the deceptive picture it could give of pupils' progress even over a very short time (a few weeks). (*Ibid.*, p. 106)

More significantly, the teachers involved reported that they had considerable difficulty in making full use of the results. As one teacher expressed it:

> In our school, standards are completely uneven. You can go from the model pupil to the one who is utterly at sea. The problem for the maths teacher is to practise differential management, to navigate at several speeds . . . The gap between pupils keeps on widening whereas we should be able to narrow it. (*Ibid.*, p. 111)

The problem of providing remedial work that was suitable for this wide range of needs was very hard to solve within the established programmes and practices for classroom teaching. The evaluation report suggested that these hopes were beyond the bounds of the possible, with some teachers saying that the goal of moving from the test results to adequate remediation would be impossible to achieve.

Other important lessons illustrated by this initiative in France are taken up below.

Formative Models of Learning

Formative assessment has to be based upon a theory about the ways in which pupils learn the subject involved. In the French example, such theory was used at the outset in the formulation of the tasks to be used. The theory must be further invoked in the interpretation of any pupil's performance, for this can also require expert judgment. Here, some help may be gained from the use of general models of the development of reasoning skills. Stiggins (1994, Chapter 10) sets out accounts of the Bloom model (already mentioned in relation to analysis of cognitive demand in domain referencing in Chapter 5) and of two alternatives — suggesting that teachers should choose whichever seems more suitable for their work. It is arguable whether such general outlines can help. There may here be a difference in practice between primary teachers, who more readily think in terms of general development, and secondary teachers who focus more on the achievements specific to their subjects (Darling-Hammond *et al.*, 1995, p. 267).

The task of analysing pupils' responses in relation to some overall scheme of progress in their learning work is a demanding one. Some of the aspects involved have already been discussed in the section on 'Aggregation, profiles and progress' in Chapter 5. The work can be helped by frameworks which have been designed and refined through trial in schools. A notable example is the Primary Language record developed by the former Inner London Education Authority, and still in use in several countries (*Ibid.*, Chapter 5). This was a scheme to be used in dialogue between the child, the parents and the school, and involved a sequence of activity starting with interviews at entry to the school, first with a child's parents and then with the child. Systematic observation of the child's language use was supplemented during the year with further conferences with the family and with the child. An end-of-year summation exercise then formulated a report for the guidance of next year's teacher. The data so obtained were interpreted in terms of reading scales which could chart progress of the child from dependence to independence and from inexperience to experience in reading. Teachers using the approach reported great surprise at the relevant knowledge that parents possessed and could convey about their children's reading. A researcher summing up an evaluation of its use in a US school commented:

> It offers a holistic framework for observing and documenting the growth of children that allows for differences in teachers as well as in children. It enables teachers to understand better how children learn and thus to teach in child-centred ways. By focusing on children's strengths, by looking at them individually, by celebrating their diversity, it supports the overall quality of school instruction. (*Ibid.*, p. 195)

This seems to show that the scheme has all the essential requirements for good formative assessment, and it is unusual for any scheme to transport so successfully from one country to another. Many frameworks which set

out criteria for progression can be found. Stiggins sets out several in his Chapter 11, the work by Ross *et al.* (1993) describes the development of a scheme for the arts, schemes for investigative science have been published by Brown and Shavelson (1996) and by Jones *et al.* (1992, Chapter 7). The schemes described in Chapter 6 for the graded assessment developments also have a great deal to offer. Where national or state curricula have been set out in levels of progression, as in the UK, Australia and New Zealand, these could also be useful, although, unlike most of the other examples, these have been developed by a bureaucratic process rather than by trial and modification in a range of classrooms.

The extent to which approaches can be general across different subjects has to be limited. Many believe that mathematics has to be built up hierarchically so that a topic must be mastered before moving on. In other areas, there is no such hierarchy, but rather a progress in sensitivity, in depth of critical judgment, and in power to articulate — as in study of literature or in the arts (Ross *et al.*, 1993) or in the three dimensional model of Christie for reading. Science might be different again, having well structured theoretical concepts, but also requiring a more complex and holistic approach if investigative experimental work is being pursued.

A development project for arts assessment found that the understanding of a model for the learning on the subjects was a major difficulty:

> Looking back at our early interviews with teachers, we can now redefine their problems with assessment as stemming from a fundamental difficulty as to how to reconcile the so-called 'teachable' elements of the arts curriculum with what is essentially 'untaught': conventional forms and expressive materials transmitted by subjective feelings. Having little concern for what Louis Arnaud Reid called 'embodied meaning', leaving little scope for a personal sense of form and feeling as an integrated whole — in other words with an inadequately defined knowledge base — a comprehensive, aesthetic assessment from our teachers was impossible. So they came across as uncertainly objective, apologetically subjective, and frequently as either radically blocking or simply ignoring the voice of the pupil. (*Ibid.*, p. 166)

The way of thinking revealed in this passage emphasizes the holistic, in part subjective, and context bound approach to assessment, in which it is seen as a qualitative art calling for the judgment of the connoisseur. This is worlds apart from the algorithmic, quantitative and objective paradigm that tends to persist in and through the psychometrics of testing. In part the polarization could be explained by the differences between (say) mathematics and art appreciation. However, where pupils' open-ended investigations in mathematics are to be assessed, there is pressure to move away from the algorithmic and quantitative paradigm. The emphasis by Popham on verbal description and judgment in the application of domain referencing, quoted in Chapter 4, reflects a similar shift.

Insofar as assessment can be diagnostic, the instruments ought to be fully researched and validated in relation to a learning model. Detailed exploration of pupils' progress in learning concepts shows that suitable models have to be multi-dimensional, even in the rather ascetic spheres of mathematics and science, so that here again there is no substitute for expert judgment of the particular situation of a pupil (Denvir and Brown, 1987; Simon *et al.*, 1994 and 1995; Brown *et al.*, 1995). The fact that very few small scale domains have been researched in adequate detail leads Wood (1991), in a review of this aspect of assessment, to be sceptical of any claims about diagnostic assessment, and to be critical of many schemes of progression that he sees as simplistic.

Whatever the model, it is essential to identify the student's level of conceptual comprehension and technical or methodological competence. The information collected, in homework and through feedback in class discussion, should be about *understanding* as well as about *subject knowledge* — it is more easy to ask questions and set exercises which call only for knowledge, and these, being more easy than others, may produce less trouble for all concerned. The research evidence is that where understanding is emphasized, even the recall of facts is more effective if pupils are tested more than a few weeks after the teaching (Nuthall and Alton-Lee, 1995).

Present Practice

Several common features emerge from surveys of research into formative assessment in many countries (Black, 1993(a); Black and Wiliam, 1997; Crooks, 1988). The overall picture is one of weak practice — formative assessment has a low rating in teachers' training and in their priorities. Key weaknesses are:

- Classroom evaluation practices generally encourage superficial and rote learning, concentrating on recall of isolated details, usually items of knowledge which pupils soon forget.
- Teachers do not generally review the assessment questions that they use and do not discuss them critically with peers, so there is little reflection on what is being assessed.
- The grading function is over-emphasized and the learning function under-emphasized.
- There is a tendency to use a normative rather than a criterion approach, which emphasizes competition between pupils rather than personal improvement of each. The evidence is that with such practices the effect of feedback is to teach the weaker pupils that they lack ability, so that they are de-motivated and lose confidence in their own capacity to learn.

Where extra emphasis on teacher assessment has been prescribed, as with the National Curriculum in England and Wales, the requirements have been widely misunderstood. Research evaluations have established that most teachers, particularly those in primary schools, have interpreted teacher assessment as summative assessment only. Some have collected voluminous records of pupils' work whilst having no clear strategy for using these records (McCallum *et al.*, 1993; 1995). Others have responded to calls for enhanced formative assessment by setting formal summative tests which mimic the external statutory tests. Recent surveys have also reported that there was very little formative assessment to be seen in science work (Russell *et al.*, 1995).

One outstanding reason for this weakness is that summative assessments, notably external testing, often dominate teaching because their results command greater esteem than those of other forms of assessment. Harry Black (1986, p. 9) has argued that it was the growth of external and summative examining that snuffed out the early development of formative assessment. However, there are other reasons for the weak development of the practices of formative assessment. They are to do with the many practical difficulties of collecting and recording evidence in the midst of all the other demands of everyday teaching, and with the challenges presented by the prospect of amending, or repeating, or differentiating teaching to respond to assessment evidence.

In a classroom, teachers are likely to look for feedback to confirm the success of their own performance. Their priority may be to keep the process going and the reaction of pupils is used to judge whether it is feasible to carry on. Teachers often choose a sub-group of only a few pupils, and it is their reactions and responses to questions which serve to justify proceeding. In an atmosphere where a fast pace is judged to be necessary, and particularly where a teacher is insecure, such practices serve the teachers' own purposes (Airasian, 1991, Chapter 4). The 'data' so collected may say little about the real learning of pupils. The same is often true of the practices used in setting and marking homework and written work in the classroom — an average mark of seven out of ten will be regarded as a satisfactory basis for carrying on even though it may mean, insofar as it conveys any clear meaning, that a significant fraction of the class may have grasped less than half of what was being taught.

Introduction of formative assessment in a thorough going way with a group of teachers calls for changes in their practice which are radical in scope and nature. Accounts of such changes have been produced, in work with groups of teachers in Australia (Torrie, 1989; Butler, 1995), and in England (Fairbrother, 1995). Some idea of the scope of the changes that might be entailed is illustrated by Torrie's study (1989) of work with Australian teachers and by study of the changes made in Queensland (Butler, 1995). In this latter innovation, the state abolished external tests and moved to sole

reliance on school-based certification. This had little effect at first on classroom practice and it required further state initiatives to move practice away from reliance on stereotyped terminal tests towards a criterion referenced basis. This latter shift was linked to assumption by teachers of enhanced responsibility for generating their own science curricula within very broad state guidelines.

Formative assessment can be hampered by misunderstandings amongst teachers. In the UK, teachers were not clear about the purpose of the teacher assessment component of the new national assessments introduced after 1988: the natural tendency was to see teacher assessment as an exercise in predicting an examination result. Given this view, the best data that teachers could collect were the results of similar tests. The use of these to generate the results throws away all the potential strengths of teacher assessment which inhere in its being based on evidence very different from that which formal testing can elicit.

Methods

Collecting Information

Most teachers have always used a variety of sources in an informal way — it is essential to sharpen this practice with a view to eliciting more usable data. Written work by pupils, both as homework or as classroom work, can only be useful if it is both designed to reveal particular features of importance, evaluated with those features in mind, and marked in such a way as to convey formative messages to the pupil.

More tightly structured information can be useful. For example, a worksheet to be completed by each student can help if the responses evoked are distinctive in that they produce detailed information in relation to statements about specific aims — i.e. they are grounded in the criterion referencing that formative assessment inevitably needs (Parkin and Richards, 1995). Such written evidence, if produced in a systematic way can relieve the teacher from the pressure of noting and recording entirely from the ephemeral evidence of classroom events. Ephemeral evidence can however have its own unique value: some teachers have found it surprisingly useful if they suspend their active teaching interventions for a time — making clear to a class what they are doing and why — and concentrate only on looking and listening with a few students at a time whilst the rest are engaged in individual or small-group activities (Cavendish *et al.*, 1990; Connor, 1990). It has also been found that teachers' learn more from looking at a few pupils in detail than in attempting to strike a common level for all — although the few have to be carefully chosen with the choice changed on each occasion (Darling-Hammond *et al.*, 1995, p. 254). There is general agreement that

observing and recording observations cannot be valuable unless significant periods of time are reserved for the purpose.

The feedback obtained in dialogue can clearly be important — but its value depends on the quality of the questioning and on the care with which responses are treated. Whereas a closed question merely leaves a pupil calculating whether to take the risk on her chances of knowing the right answer, an open one can encourage self-expression and, if responses are continually treated with care, can show the pupils that the teacher is interested in their ideas. The aim should be to establish forms of classroom discourse which encourage pupils to speak honestly, where they express their thought processes knowing that these will be taken seriously and challenged to develop their thinking. Such discussion has to be so conducted that all of a class and not just a few are taking part — techniques such as organized debates, pre-specified reporting tasks on a rota, and group discussions followed by reports can all help here (Edwards and Mercer, 1989; Stiggins, 1994, Chapter 9).

Marking

For formative purposes, it will often be particular strengths or weaknesses that might be selected for feedback discussion, even though for summative purposes, an overall holistic judgment might be appropriate, if only because each analysed atomistic component cannot be judged outside the whole context in which it functions. Where the learning work is a production task rather than a responsive task — be it an investigative or a constructive project, or an essay based on library research, or an imaginative work, the choice between holistic and analytic approaches will arise. The report by Brown and Shavelson (1996) includes a discussion of three different ways of scoring the performance on open-ended science tasks — two being analytic and the third holistic. The work by Ross and colleagues on assessment in the arts leans heavily towards the holistic approach. However the two approaches are used, it seems clear that clear qualitative descriptions of the key characteristics of a piece of work are essential for both purposes, whereas marks or grades will convey little unless, in summative use, they represent stages on a sequence of generally agreed qualitative descriptions. This atomistic–holistic debate also bears upon the validity of pupils' performances. It is well established that a pupil's performance on a test of a specific skill will not serve as a guide to the capacity to select and deploy that skill in the context of a realistically complex task. Thus atomized, out-of-context assessment is of only limited value (Black, 1990).

The collection of samples of pupils' work in portfolios may also be a helpful approach, although these are more relevant to summative use, the formative aspect residing in the immediate appraisal of the individual pieces

of work as they are produced, in the work of selecting pieces of work to present in a portfolio and in the preparation of a summary introduction to the portfolio, where those are undertaken in negotiation between teacher and pupil.

A particular feature of the ILEA language record, described earlier in this chapter, deserves attention at this point. Emphasis is placed on assessment of a pupil's attainment at the outset of schooling, which recognizes that pupils do not at any stage, even when first entering school, come to a learning programme without a personal portfolio of strengths and weaknesses, insights and blind spots, that will affect the new learning. The mathematics assessment in France described above also addressed this problem. Some such assessment appraisal on entry would seem to be an essential feature of any formative assessment programme wherever the quality and uniformity of assessments of previous work cannot be guaranteed. Airasian's (1991) book pays particular attention to this feature (Chapter 2).

Those who have carried out development work with schools focused on improving classroom assessment invariably lay stress on the power of collaborative work between teachers. For example, the summary of a set of such case studies in US schools (Darling-Hammond *et al.*, 1995, p. 253) emphasizes that through such discussions, based on sharing ideas in their reviews of the evidence of pupils' work, teachers can build up a mutual expertise that no-one else possesses and that no-one from outside could give them. This is described as the creation of communities of practice, and it highlights the centrality of assessment discussions in the professional development of teachers.

Bias

One important danger in formative — or summative — assessment by teachers is the threat of bias. Research in this area is reviewed in Wood's book and a full discussion has been given here in Chapter 4. There are conflicting results on whether teachers confuse or conflate industry and effort with achievement. There is evidence of teachers behaving differently towards boys and girls, towards pupils from different social classes, and towards good-looking and plain-looking pupils. In each of these cases, some teachers would rate a particular piece of work more highly if it came from one type of pupil rather than the other. There is also concern over some evidence that teachers assessments tend to be influenced by opinions about a pupil's 'ability' rather than strictly on achievement — it being suggested that teachers will tend to explain the differential responses of pupils to their efforts to teach as inevitable consequences of a predetermined 'ability'.

Acting on the Evidence

All collection of formative evidence must be guided by a strategy for ensuing action. The distinctive feature of formative assessment is that the information is used to modify the learning programme in order to make it more effective — there is little point in gathering information unless it can be acted upon. Since assessment information is sure to reveal heterogeneity in learning needs, the action has to include some form of differentiated teaching. One reason why this path is not followed was suggested by Perrenoud (1991) 'There is a desire *in everyone not to know about things about which one can do nothing*' (p. 89 author's emphasis).

The action which should follow any assessment revealing shortcomings in knowledge or understanding might have to be repetition — of the same work or of other work with the same aim — in order to ensure mastery before proceeding. This may be seen as essential in a hierarchical subject, but over-rigid and impracticable in other areas. A whole set of initiatives calling for quite radically new forms of classroom teaching was developed in the USA under the banner of mastery learning, a system in which each pupil would only move from one issue to the next when he had shown mastery of it, and would be required to repeat work after failures until success had been achieved. Significant gains have been reported, but the movement did not take a firm hold (Block *et al.*, 1989). Work in Scotland to establish formative and criterion-referenced assessment reported radical changes in the teachers' perceptions of their pupils and was seen by the authors as adapting elements of the mastery learning concept to the environment of classrooms in Scotland (Black and Dockrell, 1984, p. 56).

It is hardly possible to implant formative assessment into an existing teaching programme without changing that programme. New assessment cannot just be attached to an existing teaching scheme, it has to be built into a revision of the scheme, if only because its use to guide pupils' learning according to their different needs can only happen if the teaching plans allow the extensive time for planning and organization that such use requires: opportunities have to be created for making use of the feedback information provided by the assessment.

One approach uses large scale adaptations by placing pupils in separate streams, or sets, or tracks, but these do not deal with immediate needs. Some teachers have responded by organizing units of work into a core and extension, with the extension work varying widely, from advanced new topics for high attainers, to repetition of the basics for those in serious need (Black, 1995, p. 139). Others indicate less formal and more flexible approaches, building in revision or revisiting a problem through opportunities in later work for those in need. Affecting this last issue is the extent to which teaching programmes are flexible rather than rigid.

Effective use of assessment feedback requires teacher judgment, and the confidence and flexibility in management of a curriculum plan that can only come from ownership of such a plan. Thus it seems that, ideally, any scheme for incorporating good formative opportunities has to be constructed by teachers for themselves. In such construction, teachers' have to handle two innovations — the need to implement new methods for differentiation and flexibility in learning and the need to learn, perhaps invent, a new technology for producing the appropriate evidence of pupils' achievements.

The use of formative assessment feedback may be on an individual basis or at a more general level. Given that there is evidence that cooperative learning in groups can help motivation, inter-personal skills and learning, feedback which evaluates the work of a group overall can be used — and there is evidence that this can be effective if it is also combined with some evaluation of particular individual difficulties (Crooks, 1988).

Summative Assessment by Teachers

Collection and Selection

Some of the issues here have already been explored in Chapter 3, both in the section on 'Certification, progress and transfer' and particularly in the last section on 'Interaction of purposes'. Formative and summative functions can be regarded as two ends of a spectrum. At the formative end, the teacher conducts the assessment, makes the inferences, plans any action and will see and have to deal with the consequences, which should be short-term. At the summative end, the conduct of the assessment, drawing of inferences and the planning of action go beyond the teacher and perhaps the school. The teacher may have the whole, a part, or none of the responsibility and the consequences will arise in the medium to long term future. An end-of-year test conducted in a secondary school by a group of teachers and used both as input to reports to parents and as a basis for assignment to teaching groups in the next school year clearly lies between these extremes.

Many assessments carried out by teachers are in fact summative simply because they are not used to guide immediate changes in the teaching and learning work. This can be true even in relation to frequent homework and weekly tests, where, in the absence of effective feedback, the practice can be described as frequent summative. There can be no clear rule as to whether given assessment exercises are in themselves formative or summative — it will depend on how they relate to the pupils' work and to the way the results are interpreted and used. Some may be more useful for one purpose than the other, some may be useful for both.

A clearer grasp of the issues here can only be reached by exploring the purposes for which a teacher may need to produce summative assessments. Three possibilities can illustrate this point:

- If the purpose is as an overall guide for anyone in charge of subsequent study within the same general programme — for example, moving from year 8 to year 9 in the same subject department — then a profile of performance, showing strengths and weaknesses in relation to significant criteria in the subject department's common programme of work, would be appropriate. The information might best be detailed, linked to the latest achievement of a pupil in each of the areas concerned rather than to an average over time, and this would call for little or no aggregation over components.
- If the purpose is an overall guide for a new level of study in the same subject — for example, moving from a secondary school to a sixth-form college to start A-level work, then a profile of a few components (as in the attainment targets of the National Curriculum) and a summation over the last year of work might be the best guide to potential for further study in the subject.
- If the purpose is a general guide to study in a quite new field, or for employment where no specific component of the learning is directly relevant, then a summation over all components and over the last year might be best.

Any of these proposals could be implemented by using all of, a selection of, or none of the evidence gained in formative assessment work, and this source could be supplemented or replaced by results of summative tests set by the teacher. The process of selection and aggregation of the formative results would be a reinterpretation in relation to the intended purpose. Aggregation can present particular difficulties. It might seem easy to add numbers, but if these are not on a common scale the result will be meaningless. Aggregation can only be sensible in terms of a qualitative descriptive scale to which the component parts contribute in ways that are clearly documented.

Given that for summative results, with their medium to long term effects, mistakes are harder to rectify, the reliability of teachers' summative assessment must be questioned. A survey of the relevant research literature shows that assessment by teachers can attain reliabilities comparable with those of external testing, although there are large differences between teachers in this respect (Black, 1993(a), pp. 53–6). For example, with US science teachers assessing for the first time open-ended practical investigations, a programme of careful training led to their achieving a satisfactory level of reliability (Baxter and Shavelson, 1994).

Moderation and Reliability

The reliability of teachers' assessment may be inadequate because of internal errors in their procedures, or because of inconsistency in criteria and standards

between different teachers. Special methods for moderation are needed to achieve and ensure consistency between different teachers, both in the use of criteria and in the application of standards. One method is by teacher meetings, where a group, either within a school or between schools, exchange samples of pupils' work and discuss their assessments in relation to these. Such meetings have been found to be very valuable for the development of the teachers involved, often revealing how isolated teachers have been in respect of their standards and expectations. The points made in the previous section on 'communities of practice' are relevant here.

However, this process is expensive in requiring the meeting time of teachers, careful preparation and skilled leadership. Visit by inspectors, who can examine any practical or art-work and talk with both pupils and their teachers, are a useful alternative, although less flexible because the inspectors are constrained to be consistent across all visits, so that they cannot adjust as they learn from successive inspections. A third, and least expensive, method uses a common written test to determine a mean and spread for each teaching group, which is then used to scale the teacher's results for that group. This is economical, but involves the assumption that the group's attainment on the test will be a valid indicator of their attainment on the teacher assessed work.

All of these methods have been subject to research scrutiny in studies in the UK (Gipps, 1994, Chapter 4). The development of moderation methods was strongly supported in the growth of the Certificate of Secondary Education in the 1960s and 1970s, and went further with the preparatory studies for the establishing of the GCSE by 1988. A new slant was given by the growth of the several Graded Assessment schemes and Modular Schemes as described in Chapter 6. More recently, there have been enforced restrictions on this growth. The history of the use of moderation methods in the national assessments at ages 7, 11 and 14 is a complex and rather turbulent story which has been fully chronicled by several authors (James, 1994; Gipps, 1994, Chapters 6 and 7; Daugherty, 1995). In the broader context of 'quality assurance' for assessment results, it can be seen that many countries struggle with similar problems (Broadfoot, 1994). These vagaries raise the question of whether there are fundamental reasons why teacher assessment cannot serve a summative purpose. Harry Black (1994) describes how assessment in further education in Scotland has been a working success. The quality assurance mechanisms put in place using general descriptors as documented guides to standards, internal assurance procedures, and audit by external examiners, have established a common set of standards across the colleges, even though many of the those colleges involved have suspicions that uniformity is far from being achieved. The history of the assessment and curriculum development in Queensland is also a story of success — achieved after a long process of teachers' professional development (Butler, 1995).

Two Roles in Tension?

Insofar as teachers have to be involved in both formative and summative assessment, they are bound to experience some tension between their two roles. The former is private, focused on the learning needs of their pupils, which it must be their first concern to serve. The other has to reckon with pressures and constraints that come from outside. National and regional high-stakes systems create pressures for teachers to work within a framework which drives both their school policies and parental expectations, and the personal concern for the best interests of their pupils might have to take second place. The teacher has to hold the boundary between the different requirements of the two roles.

There is ample evidence of confusion arising from the overlap between these two roles. Attempts to enhance teacher assessment can too easily reduce in practice to greater use of that assessment for summative purposes, and to more frequent application of teacher assessments, with collection and storage of the results becoming a burden. The summative purpose requires the cumulative addition of results over time, leading teachers to concentrate on merely recording the results — each as a final outcome, and leading pupils to believe that the only purpose of assessment of performance is to add to a final score. Teachers can be tempted to formalize their assessment work so that they isolate it from learning development — indeed some think of assessment only in terms of use of standardized tests (Black, 1993(a), p. 52). Studies of assessment practices adopted by primary teachers in Britain following the introduction of the national curriculum and assessments gave evidence of all of these difficulties. Told that they were expected to produce teacher assessments, many responded by setting summative and terminal tests which were their imitation of the national assessment tests (McCallum *et al.*, 1993).

The traditional dominance of the summative function means that formative assessment struggles for its status and development. The summative function can inhibit the growth of the formative function because external tests are accepted as the model for teachers' assessments so driving these towards techniques appropriate only for summative use. External tests are a poor model for formative assessment because:

- In summative testing the need for a single overall result means that quite disparate data (for example, for practical and for theory) have to be added in ways that are often arbitrary: formative assessment does not have to do this.
- Summative assessment has peculiar problems with criterion referencing, partly because of the need to aggregate, partly because it cannot rely on personal judgments in deciding about the application of broad criteria to the work of individual pupils; such problems are far less serious in the practice of formative assessment.

- Summative work has to insist on standards of uniformity and reliability in collection and recording of data which are not needed in formative work and which inhibit the freedom and attention to individual needs that formative work requires.
- Whilst summative processes have to be seen to be 'fair' in treating all pupils in the same way, formative practice, with its priorities of identifying and helping to meet the learning needs of each pupil, can treat different pupils very differently.
- Summative tests are designed to give an overall score, so they work at low resolution and may yield little of the meaningful detail that formative feedback requires.
- Summative purposes can demand documented evidence for results — for example, for any auditing review — and so add to workload and distort formative practice, whereas formative work calls for action on the data rather than storage of it.

One way out of the dilemmas thus posed is to separate the two functions entirely, and some have recommended this, both on theoretical grounds that separate purposes require separate instruments and procedures, and on the grounds of the dilemmas for the teacher in being both friend and judge of the pupil (Harlen *et al.*, 1993; Black, 1993b).

The arguments could be separated into two components. The first is about the relation between the formative and summative roles that any teacher must adopt because the school itself has to generate summative results for its own internal purposes. Here it seems unrealistic to think, in the light of the discussion above about the range of summative assessments that teachers must make, that they can avoid all responsibility for summative work and the discussion has to be about how best to interrelate the generation and interpretation of data for the two roles, whilst keeping their separate purposes distinct. A key feature here is that it is the interpretation of the data, rather than the data as such, which might distinguish formative from summative practice (Wiliam and Black, 1996).

The second argument is about the tension between the roles within school and possible roles for public certification and accountability. Here, a separation might be attractive but has the following disadvantages:

- Some of the most important aims of education cannot be reflected in, and so supported by, assessment systems which rely only on short external testing.

Examples are practical work in science and technology, project work involving literature research or collection of data in the local community or environment, and performance in the arts. Teacher assessed summative components are essential to securing adequate validity in respect of these aspects. These

necessities are recognized in the national assessments, and in all the public certificate examinations in the UK. However, to use teacher assessment for these components only tends to lower their status, the implication being that teacher assessment is a second best which is only adopted where there is no better alternative.

- External tests which are economical are bound to take only a short time.

Therefore their reliability and validity are bound to be severely constrained: they can only use a limited range of methods, and must be limited in respect of their sampling of the relevant domains. It is at least arguable that both the reliability and validity in all areas to be assessed could be enhanced by a combination of teachers' assessments and external tests.

- The status of teachers is lowered if they play no part in high-stakes testing.

As long as external tests stand for the most important functions — the high stakes of certification and accountability — teachers' assessments, both formative and summative, will be overshadowed and will not receive the priority and status that they need if they are to make their full contribution to improving learning.

None of this is to argue that the two functions are the same, nor to argue that in being responsible for both, teachers do not have to distinguish carefully between them. However, a workable relationship demands more than skill and goodwill on the part of teachers — it will also be critically dependent on the context of public policy in which they have to operate. This issue will be taken further in the discussion of systemic perspectives in Chapter 9.

Developing Good Practice

The overall weight of research evidence is that formative assessment by teachers can raise standards of learning and can improve motivation provided that it meets certain conditions (Crooks, 1988). It has to be criterion referenced rather than normative, and it must be tuned to the needs of each pupil. Blame should not be used, and praise seems to be most helpful when it is used sparingly and linked with objective advice about specific performance. The feedback has to concentrate on the most important learning goals, it has to be related to a clear model of learning progress in the subject, and it has to set high but attainable targets.

Surveys of the practice of formative assessment in several countries show that its potential to improve teaching and learning has not been

exploited. Thus, this field of assessment offers distinct possibilities for the improvement of the standards of learning. However, these cannot be realized by change of assessment alone. The changes have to be part of a coherent programme in which curriculum and pedagogy are also restructured, and the tensions with summative and high-stakes factors have to be anticipated and minimized.

Overall, such visions point to a very demanding programme of development. Formative feedback implies more than correction of errors. It really needs understanding of their causes, which would tax the best learning experts. Furthermore, the student's main resource in self-adjustment is the model provided by the teacher. Students must get into the frame of reference of the teacher, to share the model of learning which gives meaning to the criteria that are reflected in assessment. Thus the development of formative assessment implies changes in role for both teachers and students (Perrenoud, 1991; Bonniol, 1991). For this reason, the discussion is incomplete without consideration of the pupils' role in assessment, which is the focus of Chapter 8.

Summary

- Formative assessment is an essential part of effective teaching and learning.
- The validity of formative assessments depends on the validity of the models of learning on which they are based.
- Current practice in formative assessment suffers from many weaknesses which might only be overcome by sustained programmes of professional development.
- Teachers can collect a wide variety of types of evidence, but these must be designed and assessed in relation to learning purposes, and with care to avoid bias.
- Teachers have also to produce summative assessments for a range of purposes, and the quality of this work and its comparability across schools can both be improved by various methods of moderation, particularly peer-group moderation.
- Summative assessment requirements readily dominate formative work because of external pressures, and will then impoverish it because formative practices provide poor models for formative work.

Bibliography

AIRASIAN, P.W. (1991) *Classroom Assessment*, New York: McGraw Hill. Chapters 2, 3, 4, 8 and 10 are all relevant.

BAIRD, J.R. and NORTHFIELD, J.R. (eds) (1992) *Learning from the PEEL Experience*, Melbourne: Monash University.

BLACK, H.D. and DOCKRELL, W.B. (1984) *Criterion Referenced Assessment in the Classroom*, Edinburgh: Scottish Council for Research in Education.

BLACK, P.J. (1993) 'Formative and summative assessment by teachers', *Studies in Science Education*, **21**, pp. 49–97.

CROOKS, T.J. (1988) 'The impact of classroom evaluation practices on students', *Review of Educational Research*, **58**, 4, pp. 438–81.

DARLING-HAMMOND, L., ANCESS, J. and FALK, B. (1995) *Authentic Assessment in Action Studies of Schools and Students at Work*, New York: Teachers' College Press. A study based upon detailed case-studies of innovations in teachers' assessments in five schools in or near New York.

FAIRBROTHER, R., BLACK, P.J. and GILL, P. (eds) (1995) *Teachers Assessing Pupils: Lessons from Science Classrooms*, Hatfield UK: Association for Science Education.

GIPPS, C.V. (1994) *Beyond Testing: Towards a Theory of Educational Assessment*, London: Falmer Press, Chapters 4, 6 and 7.

GIPPS, C. and STOBART, G. (1993) *Assessment. A Teachers' Guide to the Issues*, London: Hodder and Stoughton. Chapter 6.

PATTENDEN, J. (1995) 'Coping with assessment: An approach to recent developments (11–16)', in MARUM, E. (ed.) *Towards 2000: The Future of Childhood, Literacy and Schooling*, London: Falmer Press, pp. 148–87. Covers assessment of English, with analysis of changes since 1988, advice for departmental policies including useful samples of departmental policy and pupil assessment documents.

ROSS, M., RADNOR, H., MITCHELL, S. and BIERTON, C. (1993) *Assessing Achievement in the Arts*, Buckingham: Open University Press.

STIGGINS, R.J. (1994) *Student-Centered Classroom Assessment*, New York: Merrill/Macmillan. Chapters 10 to 15 inclusive.

TORRANCE, H. (ed.) (1995) *Evaluating Authentic Assessment*, Buckingham: Open University Press. All chapters are relevant.

WOOD, R. (1991) *Assessment and Testing*, Cambridge: Cambridge University Press. Chapters 6 and 8.

References

AIRASIAN, P.W. (1991) *Classroom Assessment*, New York: McGraw Hill.

BAXTER, G.P. and SHAVELSON, R.J. (1994) 'Science performance assessment results: Benchmarks and surrogates', *International Journal of Educational Research*, **21**, 3, pp. 279–98.

BLACK, H. (1986) 'Assessment for learning', in NUTTALL, D.L. (ed.) *Assessing Educational Achievement*, London: Falmer Press, pp. 7–18.

BLACK, H. (1994) 'The quality of assessment in further education in Scotland', in HARLEN, W. (ed.) *Enhancing Quality in Assessment*, London: Paul Chapman, pp. 87–99.

BLACK, H.D. and DOCKRELL, W.B. (1984) *Criterion Referenced Assessment in the Classroom*, Edinburgh: Scottish Council for Research in Education.

BLACK, P.J. (1990) 'APU science — the past and the future', *School Science Review*, **72**, 258, pp. 13–28.

BLACK, P.J. (1993(a)) 'Formative and summative assessment by teachers', *Studies in Science Education*, **21**, pp. 49–97.

BLACK, P.J. (1993(b)) 'Assessment policy and public confidence: Comments on the BERA Policy Task Group's article "Assessment and the Improvement of Education" ', *The Curriculum Journal*, **4**, 3, pp. 421–27.

BLACK, P. (1995) 'Lessons in evolving good practice', in FAIRBROTHER, R., BLACK, P. and GILL, P. (eds) *Teachers Assessing Pupils: Lessons from Science Classrooms*, Hatfield: Association for Science Education, pp. 137–43.

BLACK, P. and ATKIN, J.M. (1996) *Changing the Subject: Innovations in Science, Mathematics and Technology Education*, London: Routledge for OECD.

BLACK, P. and WILIAM, D. (1997) 'Assessment and Classroom Learning', *Assessment in Education* (in press).

BLOCK, J.H., EFTHIN, H.E. and BURNS, R.B. (1989) *Building Effective Mastery Learning in Schools*, New York: Longman.

BONNIOL, J.J. (1991) 'The mechanisms regulating the learning process of pupils: Contribution to a theory of formative assessment', in WESTON, P. (ed.) *Assessment of Pupils' Achievement: Motivation and School Success*, Amsterdam: Swets and Zeitlinger, pp. 119–37.

BROADFOOT, P. (1994) 'Approaches to quality assurance and control in six countries', in HARLEN, W. (ed.) *Enhancing Quality in Assessment*, London: Paul Chapman, pp. 26–52.

BROWN, M.L., BLACK, P.J., BLONDEL, E. and SIMON, S.A. (1995) 'Progession in measuring', *Research Papers in Education*, **10**, 2, pp. 143–70 and 177–79.

BROWN, J.H. and SHAVELSON, R.J. (1996) *Assessing Hands-on Science: A Teacher's Guide to Performance Assessment*, California: Corwin/Sage.

BUTLER, J. (1995) 'Teachers judging standards in senior science subjects: Fifteen years of the Queensland experiment', *Studies in Science Education*, **26**, pp. 135–57.

CAVENDISH, S., GALTON, M., HARGREAVES, L. and HARLEN, W. (1990) *Observing Activities*, London: Paul Chapman.

CHRISTIE, T. (1995) 'Defining the reading domain', in OWEN, P. and PUMFREY, P. (eds) *Children Learning to Read. International Concerns, Volume 2: Curriculum and Assessment Issues: Messages for Teachers*, London: Falmer Press, pp. 107–20.

CONNOR, C. (1990) 'National Curriculum assessment and the primary school: Reactions and illustrations of emerging practice', *The Curriculum Journal*, **1**, 2, pp. 139–54.

CROOKS, T.J. (1988) 'The impact of classroom evaluation practices on students', *Review of Educational Research*, **58**, 4, pp. 438–81.

DARLING-HAMMOND, L., ANCESS, J. and FALK, B. (1995) *Authentic Assessment in Action: Studies of Schools and Students at Work*, New York: Teachers' College Press.

DAUGHERTY, R. (1995) *National Curriculum Assessment. A Review of Policy 1987–94*, London: Falmer Press.

DENVIR, B. and BROWN, M. (1987) 'The feasibility of class administered diagnostic assessment in primary mathematics', *Educational Research*, **29**, pp. 95–107.

EDWARDS, D. and MERCER, N. (1989) *Common Knowledge*, London: Routledge.

FAIRBROTHER, R., BLACK, P. and GILL, P. (eds) (1995) *Teachers Assessing Pupils: Lessons from Science Classrooms*, Hatfield: Association for Science Education.

GIPPS, C.V. (1994) *Beyond Testing: Towards a Theory of Educational Assessment*, London: Falmer Press.

HARLEN, W. (ed.) (1994) *Enhancing Quality in Assessment*, London: Paul Chapman.

HARLEN, W. (1995) 'Standards and science education in Scottish schools', *Studies in Science Education*, **26**, pp. 107–34.

HARLEN, W., GIPPS, C.V., BROADFOOT, P. and NUTTALL, D. (1993) 'Assessment and the improvement of education, *The Curriculum Journal*, **3**, 3, pp. 215–30.

JAMES, M. (1994) 'The experience of quality assurance at Key Stage 1', in HARLEN, W. (ed.) *Enhancing Quality in Assessment*, London: Paul Chapman, pp. 116–38.

JONES, A.T., SIMON, S.A., BLACK, P.J., FAIRBROTHER, R.W. and WATSON, J.R. (1992) *Open Work in Science: Development of Investigations in Schools*, Hatfield: Association for Science Education.

MCCALLUM, B., MCALISTER, S., BROWN, M. and GIPPS, C. (1993) 'Teacher assessment at Key Stage 1', *Research Papers in Education*, **8**, pp. 305–27.

MCCALLUM, B., MCALISTER, S., BROWN, M. and GIPPS, C. (1995) *Intuition or Evidence? Teachers and National Assessment of 7-year-olds*, Buckingham: Open University Press.

NUTHALL, G. and ALTON-LEE, A. (1995) 'Assessing classroom learning: How students use their knowledge and experience to answer classroom achievement test questions in science and social studies', *American Educational Research Journal*, **32**, 1, pp. 185–223.

PARKIN, C. and RICHARDS, N. (1995) 'Introducing formative assessment at KS2: An attempt using pupil self-assessment', FAIRBROTHER, R., BLACK, P. and GILL, P. (eds) (1995) *Teachers Assessing Pupils: Lessons from Science Classrooms*, Hatfield: Association for Science Education, pp. 13–28.

PERRENOUD, P. (1991) 'Towards a pragmatic approach to formative evaluation', in WESTON, P. (ed.) *Assessment of Pupils' Achievement: Motivation and School Success*, Amsterdam: Swets and Zeitlinger, pp. 79–101.

ROSS, M., RADNOR, H., MITCHELL, S. and BIERTON, C. (1993) *Assessing Achievement in the Arts*, Buckingham: Open University Press.

RUSSELL, T., QUALTER, A. and MCGUIGAN, L. (1995) 'Reflections on the implementation of National Curriculum Science Policy for the 5–14 age range: Findings and interpretation from a national evaluation study in England', *International Journal of Science Education*, **17**, 4, pp. 481–92.

SIMON, S., BLACK, P.J., BROWN, M. and BLONDEL, E. (1994) 'Progression in understanding the equilibrium of forces', *Research Papers in Education*, **9**, 2, pp. 249–80.

SIMON, S., BLACK, P.J., BROWN, M. and BLONDEL, E. (1995) *Forces in Balance*, Hatfield: Association for Science Education.

STIGGINS, R.J. (1994) *Student-Centred Classroom Assessment*, New York: Merrill/ Macmillan.

TORRIE, I. (1989) 'Developing achievement based assessment using grade related criteria', *Research in Science Education*, **19**, pp. 286–90.

WILIAM, D. and BLACK, P. (1996) 'Meanings and consequences: A basis for distinguishing formative and summative functions of assessment', *British Educational Research Journal*, **22**, 5, pp. 537–48.

WOOD, R. (1991) *Assessment and Testing*, Cambridge: Cambridge University Press, Chapter 8.